CHARACTERISTICS OF ENSEMBLE FORECAST SYSTEMS FOR SOUTHEASTERN SNOWFALL EVENTS

A Thesis
Presented to
The Academic Faculty

By

Robert D. Haynes

In Partial Fulfillment
Of the Requirements of the Degree
Master of Science in the
School of Earth and Atmospheric Sciences

Georgia Institute of Technology
August 2017

CHARACTERISTICS OF ENSEMBLE FORECAST SYSTEMS FOR SOUTHEASTERN SNOWFALL EVENTS

Approved by:

Dr. Robert Black, Advisor
School of Earth and Atmospheric Sciences
Georgia Institute of Technology

Mr. Steven Nelson
Science and Operations Officer
National Weather Service

Dr. Yi Deng
School of Earth and Atmospheric Sciences
Georgia Institute of Technology

Date Approved: July 26 2017

Dr. James Belanger
Senior Meteorological Scientist
The Weather Company

ACKNOWLEDGEMENTS

Thanks are in order for many people, and this does not come remotely close to naming them all nor fully imparting their significance. I would first like to express appreciation to the Georgia Institute of Technology and the School of Earth and Atmospheric Sciences. After all these years, I cannot imagine having attended anywhere else. The faculty and staff have provided an experience I am glad that I did not miss.

Among the faculty and staff, I would like to thank Dr. Violeta Toma and Dr. James Belanger, who took me on, despite not having a perfectly clear direction for my research interests. Their influence is evident throughout this work. They have both helped me professionally and personally, and their hard work and dedication is an inspiration. After they left, Dr. Robert Black helped see the thesis work through to completion, and I am grateful for his sound advice and his assistance with my research.

I would also like to give regards to the Atlanta, GA National Weather Service. I would like to specially thank Laura Belanger, Adam Baker, Steven Nelson, and George Wetzel, who acted as the main focal points for the student volunteering program while I was there, as well as Keith Stellman, the MIC who has afforded me the opportunity.

Finally, I would like to extend thanks to friends and family. My family has been supportive the entire way and was a comfort when things were down. My cousin and roommate, Austin Rutherford, whose diligence and desire to succeed and take on new challenges is also an inspiration, and the local congregation at Embry Hills, with whom I've worshipped and been edified with most of my entire life, and to the God that made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

AV	Actual Variance
CAO	Cold Air Outbreak
DJF	December, January, February
EC	European Center
ECMWF	European Center for Medium-Range Weather Forecasts
EDA	Ensemble Data Assimilation
EPS	Ensemble Prediction System
ERA	ECMWF Reanalysis
ETS	Equitable Threat Score
EV	Ensemble Variance
GA	Georgia
GEFS	Global Ensemble Forecasting System
GH	Geopotential Height
IFS	Integrated Forecast System
KATL	Atlanta Hartsfield-Jackson Airport
lat	Latitude
LLJ	Low-Level Jet
lon	Longitude
M	Total number of ensembles
m	Individual ensemble member
MOS	Model Output Statistics

mROC	Modified Relative Operating Characteristic
N	Total Number of bins
n	Individual Bin Number
NCEP	National Center for Environmental Protection
NOHRSC	National Operational Hydrologic Remote Sensing Center
NWS	National Weather Service
ob	Observed Value
QPF	Quantitative Precipitation Forecast
R^2	Coefficient of Determination
ROC	Relative Operating Characteristic
s	Snowfall Value
SE US	Southeastern United States
SKEB	Stochastic Kinetic Energy Backscatter
SLR	Snow-to-Liquid Ratio
SPPT	Stochastically Perturbed Parameterization Tendencies
TIGGE	THORPEX Interactive Grand Global Ensemble
TS	Tropical Storm
WFO	Weather Forecasting Office
WPC	Weather Prediction Center

SUMMARY

The use of ensemble forecasting has burgeoned with the advent of greater technological resources. Forecasters now have available to them a range of possible forecast outcomes that can be utilized to understand forecast biases and to convey the most likely or worst case scenarios. However, ensemble forecasting systems must be used responsibly. If an ensemble is under-dispersive or heavily-biased, the observations may fall outside the ensemble, in which case it provides little useful information. Examining the behavior of the European Center's Ensemble Prediction System (EC EPS), these characteristics are analyzed in terms of snow water liquid equivalent forecasts for the 2010-2015 DJF period. Selected for the advantages of a large member count, a linear regression model is created using bins populated by ensemble variances and error variances of the EC EPS, ordered by ascending ensemble variance, to calibrate the ensembles to provide a 1:1 prediction of the range of errors from the ensemble mean forecast. Training statistics indicate the model based on ensemble variance explains more than 90% of the error variance at all lead times. However, in-depth analysis of two out-of-sample, disparate winter weather events demonstrate how the linear regression model relies too heavily on the ensemble spread and can over-forecast or under-forecast depending on how large the ensemble spread is relative to the 2010-2015 DJF period. All things considered, the application of a linear model relating ensemble variance to error variance provides a promising means to adjust an ensemble system's bias in estimating the range of possible outcomes for snowfall in the Southeastern US (SE US).

CHAPTER 1

INTRODUCTION

Providing accurate and timely forecasts of snowfall, especially in the SE US, is challenging. Thermodynamic profiles of the lower troposphere are often simultaneously unsaturated and below-freezing in this region. Thus, the conditions required for snowfall are very restrictive, with a slight change in a synoptic weather system's intensity and location enough to create modest forecast errors. Due to this uncertainty, officials are challenged locating the best regions to issue warnings and advisories and reluctant to make changes to avoid inconsistent messaging. Among different methods of numerical weather prediction, modern forecasting research has recently focused on the utilization of ensembles forecast systems to provide probabilistic forecasts to account for such uncertainty (Gneiting and Raftery 2005). Ensemble prediction systems provide consumers with a range of solutions to account for potential errors in model initialization and model parameterizations. However, operational forecasters do not have time to analyze each individual ensemble and the combined impact on the ensemble mean. Developing an automated means of efficiently interpreting an ensemble system's output would be of great value to operational forecasters and decision makers.

Rarity of Snowfall and the Predictability Problem

The relatively low number of snowfall events in the Southeast provides some forecasting challenges, but also provides a set of advantages. Compared to other regions for the same recurrence interval, the SE US has noticeably lower snowfall thresholds (Squires et al. 2014). A recurrence interval of 25-yrs or a 4% chance of occurrence

within a 48-hr time frame ranges from less than 15 cm (~6 in) to 30 cm (~12 in) across the SE US, while most other regions are upwards of 61 cm (~24 in). Estimates across the SE US, including much of Georgia, reveal an average of about 2 snow days in a winter season for the period 1981-2010, with average snowfall amounts of around 4 cm (~1.5 in). A marginally higher frequency of snowpack per season indicates that whatever snow falls does not remain for extended periods of time. In the state of Georgia and the region covered by the Atlanta Weather Forecasting Office (WFO), snow events are typically low magnitude (Notaro et al. 2014).

However, low magnitude events tend to be operationally problematic due to the difficulty of forecasting and assigning expected impacts for each system. Different warnings, advisories, statements, and other products issued by the National Weather Service (NWS) are designed to communicate the expected impacts. Those impacted are then expected to take the necessary actions to ensure their safety, which is problematic in the instance where low-magnitude events cause impacts greater than whatever forecast product was issued (DeVoir 2004). Weaker systems are harder to forecast properly, especially winter weather systems. However, this does not diminish their importance. Slight changes in the thermodynamic profile or vertical forcing can produce vastly different outcomes, even in scenarios with similar synoptic patterns (Homan and Uccellini 1987, Lackmann et al. 2002).

While the low number of snowfall events may provide several statistical disadvantages for study, the benefit is that the variety of synoptic weather patterns capable of producing snow in the SE US is limited. Climatologically, the SE US is above freezing for much of winter (DJF). An antecedent cold air outbreak (CAO) is required to

maintain negative temperature anomalies. The CAO is often associated with anomalous surface high pressure stretching from the Northern US to the Gulf of Mexico. Furthermore, SE snow events are also associated with 500 hPa shortwave features embedded within an anomalous longwave trough positioned along the East Coast (Konrad 1996). Since CAO events bring in dry, continental polar air at the lower levels, moisture must be advected from the Gulf of Mexico or the Atlantic to produce precipitation. In heavy snowfall cases, vertical ascent through isentropic upglide of available moisture plays a significant role. This isentropic upglide is enhanced when a low-level jet (LLJ) streak allows for efficient moisture transport from the developing low pressure towards the building high pressure area (Mote et al. 1997). Latent heat processes occurring as moisture is transported along this upglide region can also enhance vertical motion or system development. However, it is also important that any warm-air advection from the Gulf of Mexico or latent heating is not so strong that it results in a warm layer aloft with temperatures above freezing. Given that these systems are often weak and/or early in their life-cycle when impacting the state of Georgia, it is more difficult for ensemble systems to consistently depict the location of cyclogenesis and the phase transition regions. This impacts how models depict warm-air advection and diabatic processes common to SE US winter weather events (Gurka et al. 1995; Keeter et al. 1995; Ebert 2001).

Given the wide range of possibilities present in a winter weather event, ensemble forecast techniques are becoming increasingly popular. For example, the European Center's Ensemble Prediction System (EC EPS) utilizes 51 non-linear simulations to produce a set of probability distributions to forecast atmospheric conditions (ECMWF

2017). Whereas a deterministic model gives users one solution, an ensemble system gives users probabilistic distributions based on multiple outcomes. This feature allows users to determine the likelihood of an event and provide an explicit statement regarding the uncertainty of a weather forecast (Palmer et al. 2000). Stating information in terms of probability allows for a more diverse means of forecast analysis, whether it be weighting schemes, bias correction techniques, or ensemble calibration (Ebert and McBride 2000; Nehrkorn et al. 2003; Keil and Craig 2007; Wilks and Hamill 2007; Du and Zhou 2011; Andersson and Tsonevsky 2015). Probabilistic measures of skill have also been developed, such as the relative operating characteristics (ROC) score, to assess the ensemble's skill via numerical yes/no decision thresholds (Harvey et al. 1992; Mason and Graham 1999).

For this analysis, the ECMWF EPS, or EC EPS, was selected due to its large ensemble size and relatively-high spatial and temporal resolution when compared to other ensemble forecast systems. However, we note that neither the uncalibrated EC EPS nor the Global Ensemble Forecasting System (GEFS) produce particularly skillful 5 mm quantitative precipitation forecasts (QPF) and that improvements to precipitation forecasts are challenging to make (Hamill et al. 2008; Haiden et al. 2015). Bias correction techniques, such as Model Output Statistics (MOS), are readily applied to single-forecast runs by operational forecasters. New techniques are constantly being developed for understanding and assimilating the information provided by the ensemble forecast system. For this analysis, the ensemble spread-error relationship is studied to improve upon the understanding of the ensemble mean forecast using the 2010-2015 winter seasons as a training period (Gneiting 2014).

Framework of the EC EPS

To better facilitate an understanding regarding ensemble forecasts in the context of this study, it is necessary to first understand the history of the EC EPS and how it is currently structured. In December 1992, the National Meteorological Center (modern day National Center for Environmental Prediction) and ECMWF began issuing operational ensemble forecasts. Tracton and Kalnay (1993) speculated on the practical applications possible with the new operational forecasting tool ranging from probabilistic forecasting based on ensemble spread and developing new means to quantify uncertainty. Initially, most research was related to optimally generating perturbed initial conditions. NCEP focused on using perturbed initial conditions without rapid error growth by rescaling with a control forecast acting to adjust for expected errors, through the method known as a "breeding of growing modes" (Toth and Kalnay 1993). The ECMWF used a singular vector approach based upon the fastest growing modes in the successive short-term forecast determined the evolution of the forecast (Palmer et al. 1992). Singular vector perturbations were completed through error estimates of the analysis. Each singular vector has a Gaussian distribution calculated to account for the analysis errors and to scale the errors for dispersion, which become the coefficients of perturbation. Error adjustments were thus restricted to operational analyses early on (Molteni et al. 1996).

Over time, model errors in the physical parameterization of certain processes and errors caused by resolution constraints were also included. The Stochastically Perturbed Parameterization Tendencies scheme (SPPT), and a Stochastic Kinetic Energy Backscatter scheme (SKEB) were developed to account for these issues (Buizza 1999;

Palmer et al. 2009; Berner et al. 2009). These changes were based on a flow-dependent group of observations. Known as the Variable Resolution Ensemble Prediction System (VarEPS), this technique altered the resolution of the ensemble forecast system to reduce computational costs of processes that no longer contributed to skillful forecasting in the long-term (Buizza et al. 2007). Later, methods following ensemble-based data assimilations (EDA) were pursued. This relatively new technique provides several advantages over VarEPS and the singular-vector approach, such as the removal of perturbed initial conditions and greater simplicity when performing the error analysis (Whitaker et al. 2008; Isaksen et al. 2010). Current EC EPS cycles use a combination of the EDA method and singular vector methods (ECMWF 2017).

While the previous discussion summarizes the evolution of forecast spread due to representation of initial conditions and parameterized perturbations, physical changes are also included within the Integrated Forecasting System (IFS), which is also the atmospheric module component for the high-resolution ECMWF. Roughly every 6 months, changes are implemented for the operational IFS ranging including improving physical processes, increasing model resolution, providing new forecast parameters, and additional observations in data assimilations, etc. For example, sleet and freezing rain parameters were not generated until May 2015 in IFS cy41r1. Prior to this model cycle, the microphysics scheme designated all non-frozen precipitation as snowfall (Forbes et al. 2014). Methods for evaluating model performance have also expanded. A reforecast, or hindcast, is run twice weekly with the latest versions of the EC EPS. The reforecast is used to derive a "model climatology" for the previous 20 years using the ECMWF Reanalysis (ERA) Interim reanalysis data for initialization. This dataset allows for the

correction of model biases and the constant reevaluation of each model cycle's performance (Hagedorn 2008). The EC EPS is a constantly evolving system improving not just the forecast, but also adding new tools for analysis in line with the latest research[1].

How Forecasters Incorporate Ensemble Forecasts

Meteorologists have two different methods of numerical weather prediction available to them: the deterministic model and the stochastic (probabilistic) model. The deterministic model generates a forecast based on the initial conditions provided to create a single forecast, while the stochastic model is generated by altering parameterization of the initial conditions or the physics of an ensemble member to generate multiple forecasts (Sivillo et al. 1997). For an operational forecaster, an ensemble forecast is one among dozens of other deterministic model forecasts that are employed. Ensembles assist in addressing issues regarding uncertainty through the spread of their various model solutions. Aside from data availability, forecasters are most often concerned with under-dispersion, in which observed outcomes lie outside the ensemble forecast range. Such problems arise in extreme or anomalous cases, with additional concerns that the ensemble means can generate atmospheric conditions that are physically inconsistent with other outputs (ECMWF User Guide). Although bias correction and weighting can be performed, biases can reverse with transitions in highly dynamic flow. The skill of an EPS is determined through both quiet and active weather patterns over an extended period of time. However, the forecaster is typically more interested in potential high-

[1] http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model: Has list of changes, updates, and revisions made to the IFS.

impact cases and, unless greater skill is demonstrated in an EPS than a deterministic forecast, will often lean towards the deterministic model or only examine the ensemble mean. Emphasis is thus placed on utilizing the EPS for providing uncertainty wording instead (Novak et al. 2008).

Evans et al. (2014) provide an interesting forecast experiment where nine forecasters examined a blinded set of QPF forecasts related to Tropical Storm Fay in 2008. Forecasters received the information as if it were a simulated forecast of a tropical cyclone along the Texas coastline and TS Fay was given a pseudonym. In the study, forecasters were given deterministic models and guidance forecasts to produce a QPF forecast. Once the forecast was returned, they were given a convective-permitting ensemble and tasked with altering their forecasts accordingly. Afterwards, each of the forecasters were given a questionnaire related to the changes they made when considering how the ensemble forecast affected their QPF predictions. Based on the numbers, the consensus forecast improved across all precipitation thresholds after consideration of the ensemble forecasts despite some discrepancies among individual forecasters. Several forecasters reported that the ensemble members depicted one of two precipitation maxes, which indicates they visually distinguished information regarding the distribution before investigating further. Overall, forecasters indicated a preference for the forecast range represented visually as opposed to analytics containing specific details about the ensemble.

Here, we are interested in characterizing the ensemble behavior of the EC EPS related to winter weather systems in the Southeast. Focusing on the 2010-2015 winter seasons, a visual range of the possibilities for singular events as well as the whole period

are examined. The primary goal is to characterize the behavior of the EC EPS and to utilize ensemble spread information to generate a linear model that quantitatively relates the ensemble variance to error variance in the forecast. This research seeks to ascertain the importance of the error-spread relationship, and whether it can be reasonably applied to address forecaster skepticism of under-dispersion in the context of snowfall forecasts. Whether forecaster skepticism of model uncertainty is warranted, this work seeks to address how forecasters can utilize the information to assess the ensemble spread of synoptic weather features through the error variance. A linear regression model is constructed with the ensemble variance as the predictor and the error variance as the predictand for analysis of the 22-23 January 2016 event and the 6-7 January 2017 event. Chapter 2 of this document provides the data sets and methods used for evaluating the EC EPS in this study. The results are discussed in Chapter 3, which begins with synoptic analysis for the 22-23 January 2016 and 6-7 January 2017 event. The primary focuses are the study of ensemble variance and error variance model in the context of how skillful the EC EPS is for the 2010-2015 DJF and the model tendencies in relation to observations through rank histograms. The results of the linear model of the ensemble variance and error variance are applied to the winter events in January 2016 and January 2017 are then presented. Finally, Chapter 4 concludes with a review of the results and the implications for EPS users for winter weather in the SE US.

CHAPTER 2

DATA AND METHODOLOGY

Datasets

The EC EPS was the primary ensemble forecast system analyzed in this study. The decision to analyze the EC EPS over NCEP's GEFS was based on previous studies suggesting the EC EPS tends to be more reliable (Buizza et al. 2005; Hamill et al. 2008) and that results for modeling error variance based on the ensemble variance is more successful with a larger ensemble family (Kolczynski et al. 2011). That is not to say the EC EPS will always produce the best forecast. Operationally, the EC EPS has 51 members that are run at 0000 UTC and 1200 UTC. The International Grand Global Ensemble (TIGGE), which collects ensemble forecasts from several numerical weather prediction models including the EC EPS, was the primary source for accessing and downloading the forecasts. A 0.25° horizontal grid-spacing in a rectangular latitude-longitude grid with the surface snowfall forecasts were extracted for DJF 2010-2015. Note that this period includes several updates to the IFS used by the EC EPS. The 0000 UTC forecasts for each day were analyzed with snowfall forecasts designated into a 24-hr period from 1200 UTC to the following day calculated up to a lead time of nine days.

The ERA-Interim reanalysis, which contains a wide range of observations from 1979-present, was utilized to verify snowfall forecasts from the ECMWF EPS, and the 1986-2015 period for the analysis of climatological anomalies for the January 2016 and January 2017 snowfall events. Similar to the EC EPS, ERA-Interim data was collected on a 0.25° horizontal grid-space for the same time period at 1200 UTC. As the name suggests, ERA-Interim is a temporary dataset that is being utilized until more advanced

methods of data assimilation are introduced with ERA-5 (Dee et al. 2011). Given the need for the reanalysis to be physically consistent, ERA-Interim is based off an earlier version of the IFS (cycle 31r2) to generate analysis and short-term forecast fields that cycled through a data assimilation system as remote and in-situ observations are ingested. It should be noted this provides a constraint, as it partly depends on a similar framework to older versions of the EC EPS, which is potentially problematic for verification of sensitive variables such as snowfall that also once included sleet and freezing rain in earlier IFS cycles. Furthermore, precipitation fields are produced through the forecast model and the assimilated information of temperature and humidity. Although ERA-Interim/LAND provides a more robust measure through the inclusion of satellite and rain gauge data, at the time of writing, the information is only available for the 1981-2010 climate period and will likely not be available till the next climate period is measured. However, the ERA-Interim still provides modest accuracy for study, despite present biases (Dee et al. 2011; Balsamo et al. 2015). For the sake of direct comparison with model forecasts, bias corrections were not computed for snowfall. This is consistent with the use of the uncalibrated EC EPS forecasts by an operational forecaster. ERA-Interim data was also used to measure the relationship of the ensemble variance and error variance and to provide verification of discrepancies in the development of the ensemble variance and error variance relationship.

Furthermore, for the sake of analyzing the January 2016 and January 2017 winter weather events, synoptic analysis for each event was assessed. Anomaly fields were calculated with the purpose of determining whether the synoptic weather pattern was consistent for typical winter weather events in the Southeast based on Konrad (1996) and

Mote (1997). The 500 hPa geopotential height (GH) and contour-shaded anomalies were utilized for the assessment of the appropriate longwave trough pattern, as well as the layer-mean temperature of 1000, 925, and 850 hPa layers to determine the presence of anomalously cold air in addition to the 1000-850 hPa thickness. Contours of the 850 hPa GH values to determine the presence of a LLJ and to indicate wind direction, assuming geostrophic flow were used to assess regions of likely temperature and moisture advection. Finally, the mean-layer specific humidity for the 1000, 925, 850, and 700 hPa layers was assessed to determine low-level moisture anomalies. These characterized regions with significant moisture and whether the events displayed typical characteristics for SE US winter weather events. Observations were averaged across a 48-hr time period through 12-hr time steps. The observational period for each snowfall event was defined as 0000 UTC 22 January 2016 to 1200 UTC 23 January 2016 and 0000 UTC 6 January 2017 to 1200 UTC 7 January 2017. The climatology was calculated with the synoptic monthly means averaged with the 0000 UTC and 1200 UTC time stamps. In order to include the recent 2010-2015 DJF period which was used for creating a linear model relating the ensemble variances and error variances, climatology was defined as 1986-2015.

Given that the ERA-Interim dataset and EC EPS are fundamentally based on the same atmospheric model, the National Operational Hydrologic Remote Sensing Center (NOHRSC)[2] was used as an additional source for verification of snowfall that was applied to the ensemble variance and error variance relationship, and for establishing the

[2] An overview of the NOHRSC dataset including relevant research, technologies, policies, and observation requirements can be found at… https://www.nohrsc.noaa.gov/technology/

modified relative operating characteristics (mROC) for the region of study. This dataset was utilized for verification because it includes information from cooperative observers and first order stations, allowing for a more substantial verification than the ERA-Interim dataset. Each observation is reported for the 24-hr period at 1200 UTC and is interpolated to a 0.25° latitude-longitude grid to match the EC EPS forecast grids from a text file containing the positions of submitted snowfall reports. Each report was required to be submitted within 3-hrs of 1200 UTC. The interpolation was performed through a Barnes objective weighting scheme similar to those used for winter weather verification by the Weather Prediction Center (WPC 2016). Three radii scans at 1.5°, 1.0°, and 0.5° were respectively weighted 0.1, 0.2, and 0.7. In other words, closer observations were weighted more for interpolating each grid space. Information from these textual reports are measurements of snowfall accumulation as opposed to liquid equivalent. Based on research from Roebber et al. 2003 and Baxter et al. 2005, a snow-liquid ratio (SLR) near the average for Peachtree City of 12:1 as opposed to the basic 10:1 was utilized. This mainly affected the slope of the relationship between ensemble variance and error variance. Given this work is interested in characterizing this relationship and discussing its potential functionality across several events, determining the SLR for each individual event was not pursued, but is recommended for future efforts seeking greater precision across singular events.

Determining Predictability of Snowfall in the EC EPS

Measuring the reliability of the EPS was computed through the mROC and provides a means to compare the impact that high predictability events and low predictability events have on modeling the error-spread relationship. This probabilistic

measure is an effective means for determining a models skill based on the false alarm ratio versus the hit rate, which is advantageous for describing sensitive forecast variable such as snowfall. For a dichotomous outcome, such as snowfall versus no snowfall, the mROC score defines the capability for the EC EPS to designate the correct outputs based on the observational dataset given. A forecast event that is successfully observed is considered a hit. A false alarm would represent a case where a forecast event was not observed. Instances where an event is observed without a forecast occurrence is considered a miss, and a forecast that appropriately forecasts no event to occur is considered a correct rejection. The hit rate describes the ensemble's capability of correctly forecasting an event compared to the total number of observed events and forecast, while the false alarm ratio is established by the number of false alarms in the forecast compared to the total number of forecasts. The score is modified including forecast versus observed snowfall at a specific threshold throughout the region of study (Peterson et al. 2008). For this study, we assume that the occurrence of any snowfall to be impactful in the SE US. Therefore, the mROC score was calculated using any snowfall. A perfect forecast correctly designates a hit or correct rejection based on the probabilities designated through the ensemble forecast. The mROC score is calculated through integrating increasing probabilities between the false alarm ratio and hit rate at specific numerical thresholds, which is anything above zero snowfall in this study. If the mROC score is greater than 0.5, then the forecast performs better than a randomly assigned forecast (Harvey et al. 1992; Andersson and Tsovenosky 2015).

Calculations of the mROC score were performed for the region designated 30°N to 40°N and 80°W to 90°W (Figure 1). This contains much of what is commonly

referred to as the "Southeastern US (SE US)", although no consensus or official definition exists for what is included in the region. For this study, the SE US is synonymous with the region of analysis. The mROC score was calculated across the period 2010-2015 DJF for this region with any snowfall as the threshold for defining a hit or a miss. The NOHRSC snowfall data was used as the observations for which the mROC score was computed. In addition to analysis across the entire time frame, the mROC score for a specific set of winter events was calculated to determine whether the absence of winter weather events during the period had an impact. Assessing the skill of the 2010-2015 DJF period and the January 2016 and January 2017 events was also used in assessing performance of ensemble variance and error variance relationships.

The winter weather events selected from the 2010-2015 DJF period were based on whether snow was reported at the KATL station records located at the Atlanta-Hartsfield Jackson Airport or if the Atlanta, GA Weather Forecasting Office (WFO) generated a summary page related to the event[3]. Table 1 lists the specific dates, presents the reason for inclusion, and notes the main region impacted relative to the state of Georgia. Winter storms within the past events page of the National Weather Service include ice storms, which were included based on freezing rain and sleet being a component of snowfall outputs in earlier EC EPS cycles for comparing their relative contributions. This permitted the analysis of active winter weather events to answer potential forecaster concerns regarding the influence of quiet winter weather and provided a framework for discussing potential reasons for an event's predictability.

[3] The Atlanta, GA WFO maintains a list of summary pages for historical weather events at the following web address: http://www.weather.gov/ffc/past_events

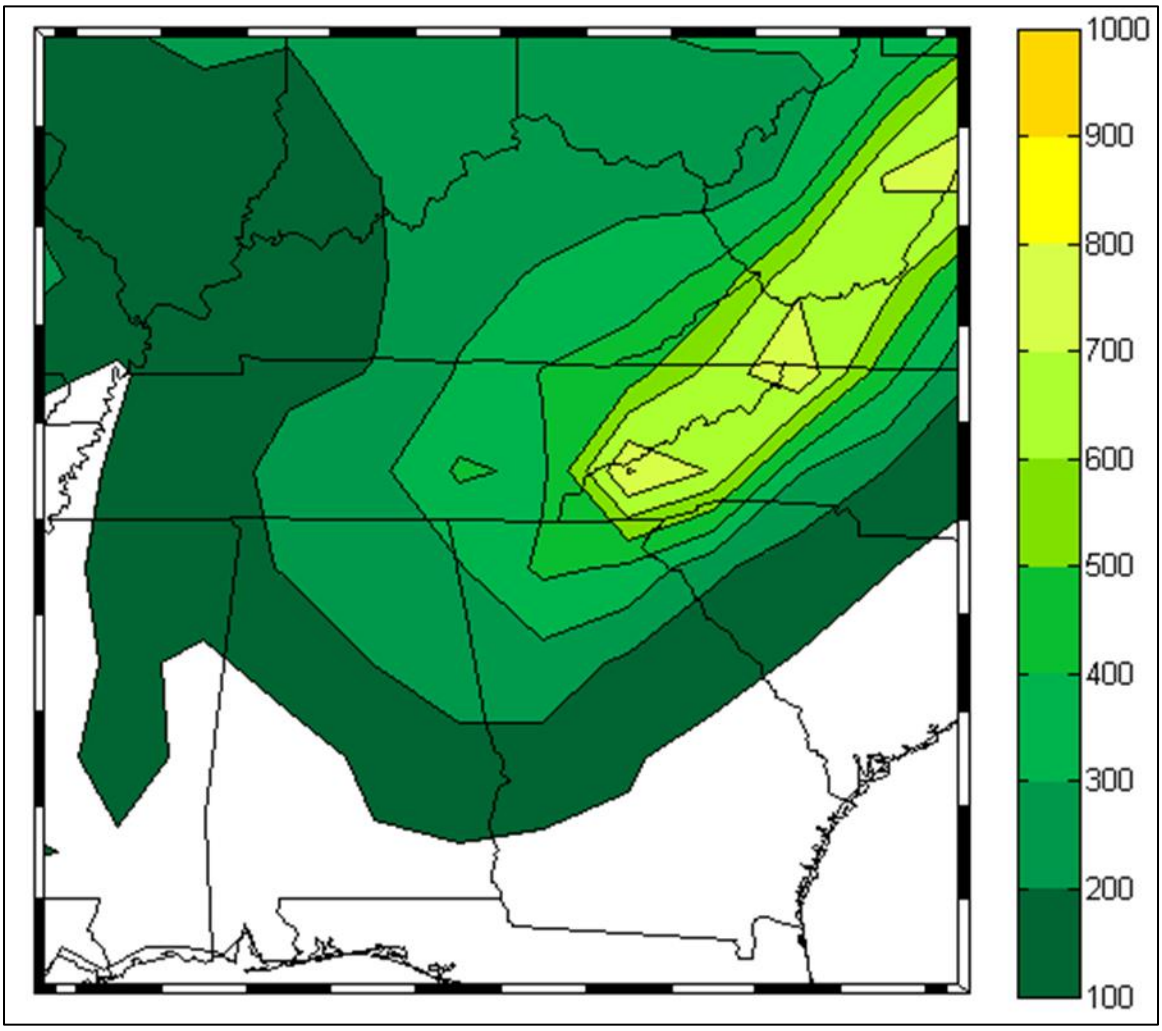

Figure 1: Region of study 30°N to 40°N and 80°W to 90°W with 1° resolution local elevation (m) imposed.

Table 1: Dates of active winter weather as noted by either KATL station records or the National Weather Service for DJF 2010-2015. The source of inclusion includes the main precipitation type. The third column notes the location of greatest snowfall in Georgia or if snow was outside Georgia for ice storms.

Date	Source of Inclusion	Region of Greatest Snowfall
2010-01-08	KATL records 1 in. snow	North Georgia
2010-02-12	KATL records 3 in. snow	Central Georgia
2010-02-16	KATL records trace snow	Georgia Mountains
2010-12-13	KATL records trace snow	Georgia Mountains
2010-12-16	NWS records ice storm	Outside Georgia
2010-12-26	NWS records widespread snow	North Georgia
2011-01-10	KATL records 4 in. snow	North Georgia
2011-02-04	NWS records wintry mix	North Georgia
2011-02-10	KATL records 1 in. snow	North Georgia
2013-02-02	NWS records light snow	Georgia Mountains
2014-01-29	KATL records 2 in. snow	Central Georgia
2014-02-13	NWS records historic ice storm	North Georgia
2015-02-15	NWS records ice storm	Outside Georgia
2015-02-21	NWS records light snow	Northwest Georgia
2015-02-25	NWS records powerful snowstorm	Northwest Georgia

In total, 15 winter events were included in determining the mROC score for active winter weather impacting Georgia. Each of these winter events were placed into one of three categories designated as high predictability, marginal predictability, or low predictability. Buizza et al. (1999) outlined a standard set of ROC scores for useful liquid equivalent precipitation forecasts, suggesting a score of 0.8 marks a successful EPS. However, given the rarity of snowfall in the southernmost extent of the analysis region and that many events contain mixtures of frozen and liquid precipitation, these events were categorized less strictly. A high predictability event was considered if the average mROC score of the first three days-lead time was at least 0.6. An event was considered marginally predictable if the average for the first 3 days-lead time was at least 0.5. Finally, a low predictability event was categorized for a mROC average well below 0.5.

Forecast Error Spread based on Ensemble Variance

With the first subsection describing the EC EPS skill and an understanding of the characteristic synoptic weather patterns associated with SE US winter weather, the second subsection is interested in analyzing the ensemble variance and describing characteristics of the ensemble forecasts from the 2010-2015 DJF period to two recent events. With the understanding that simply looking for the best ensemble members and eliminating the worst ensemble members is not valuable, an approach to assess the EPS behavior based on variance of the ensemble members and the ensemble mean from observations was sought (Bright and Nutter 2004). A method for generating a linear model of the ensemble variance and the error variance was selected (Kolczynski et al. 2011). Given the importance of the ensemble spread and the error spread to the construction of the model, snowfall output of the EC EPS members were compared to ERA-Interim observations via rank histograms for the specific winter weather events to address forecast tendencies of the EC EPS. The rank histogram compares observations to each ensemble member and designates rank of the observed value amongst the ensemble members (Wilks 2014). For M ensemble members, there are M+1 ranks. In the case of the EC EPS, there are currently 51 ensemble members. A rank of 1 indicates the observed value was lower than all members and a rank of 52 indicates the observed value was higher than all members. Ties, which were most common for regions with no snowfall, were broken using a uniform distribution of randomly generated numbers to designate the rank within the set of ensemble members that tied with the observation (Hamill and Colucci 1997). Ideally an ensemble displays rank uniformity, indicating equal probability for the ensemble members to cover the range of observational values.

Lack of rank uniformity indicates failure of initialization and EPS biases (Hamill 2001). The forecast snowfall and the observations were verified for a 48-hr time period of the snowfall event with the dates listed in Table 1, except for the February 2014 event, which had snowfall spanning across a 72-hr time period. The rank histograms were constructed using both the ERA-Interim reanalysis and NOHRSC data to demonstrate differences between the datasets. Furthermore, the events in January 2016 and January 2017 were included. The rank histograms provide a visual description of whether the EC EPS is under-dispersive and to what degree.

Several studies have sought to find a relationship between model skill defined by their error with the ensemble variance and realized the lognormal variance of the ensemble and assumed normal distribution of error yielded weak correlations (Houtemaker 1993; Grimit and Mass 2007). Thus, the linear relationship between the second moment of the ensemble distribution and the error distribution of was analyzed (Kolczynski et al. 2011). Individually, the ensemble variance and error variance are highly scattered, but when averaged into bins, a linear relationship can be established between the ensemble variance and the error variance. Following Roulston et al. (2005) and Kolcyznski et al. (2009), the ensemble variance (EV) was calculated by a sum of the square differences between each individual ensemble and the ensemble mean at each grid point.

$$EV = \frac{1}{M}\sum_{m=1}^{M}(s_{m(lon,lat)} - \overline{s_{(lon,lat)}})^2 \quad (1)$$

Thus, *EV*, which represents the ensemble variance, was calculated for snowfall (s) at each longitude and latitude point. *M* represents the number of ensemble members – 51 within the EC EPS when including the control forecast. The value for each ensemble

member, s_m, indicating the snowfall output of that given member. The error variance or actual variance, labeled AV, is generated through the mean-square error of the ensemble mean summed into bins of equal size and averaged based on the bin's size. Each bin is defined as a singular value defined by the mean of the observations placed inside. Based on the previous works mentioned above, each bin contains N equal to 1000 observations. The error variance was first sorted in the order of ascending ensemble variance before being segmented into the various bins.

$$AV_N = \frac{1}{N}\sum_{n=1}^{N}(s_{ob,n(lon,lat)} - \overline{s_{n(lon,lat)}})^2 \quad (2)$$

Where s_{ob} is the observed snowfall at each grid point in ERA-Interim, AV_N indicates the actual variance or error variance value of one bin. Assuming the error spread of each forecast to be similar between forecasts of similar ensemble spreads, Equation 2 contains as many iterations as the number of defined grid points with as many forecast runs for the 2010-2015 DJF period. The total number of AV_N values is thus determined by however many times a bin can be filled with N equal to 1000 observations with any remainder left out of the population sample. This mainly acted to remove the most extreme data-points outside of the population. Across 509 forecast days and a 41x41 rectangular grid are 855,629 observations. Thus, approximately 855 AV_N calculations were completed at each lead time. Values for which information was missing was excluded, which varied with each lead time, meaning that the actual number of AV_N calculations was different across each lead time. The sorted values of the ensemble variances when binned become

$$EV_N = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}(s_{mn} - \overline{s_n})^2 \quad (3)$$

The values EV_N and AV_N were displayed on a scatter diagram and the quality of the linear fit was calculated through the coefficient of determination, R^2. As established in Wilks (2014), R^2 is calculated as a ratio between the variation in the predictand and the regression with the ensemble variance. This value indicates how successful linear regression of the predictor would be in forecasting the predictand. Thus, a value of 1 represents a perfect predictor. The population was established under the assumption that for each day, the ensemble distribution was not largely different from one day to the next and each forecast day was treated as independent from any other forecast day. Thus, the population size includes every forecast day in the 2010-2015 DJF period, but the results were separated according to lead-time, given the growth of model errors with time. This also had the benefit of setting new coefficients of regression based on forecast lead time. As forecast error growth is non-linear, the ensemble variance and error variance relationship alters the slope of the linear model for each lead day.

This linear regression model was applied for a one-day lead time for the 22-23 January 2016 and 7-8 January 2017 events to compare model results with observed values. The model was constructed utilizing ERA-Interim reanalysis and NOHRSC datasets to note the effect of the verification networks on the linear regression model. Given that the modeling of the error variance is based upon the ensemble variance, the verification network selected will affect the slope of the linear regression model. The difference between modeled error variance and the observed error variance was calculated for the region of study using the Equitable Threat Score (ETS). This quantity also provides a measure of the success rate of the two verification datasets (Wilks 2014).

Previous studies mentioned have measured the relationship based upon synoptic weather variables such as wind or 500 hPa geopotential heights, whereas snow is a quantity produced within the model's micro-physics schemes that combine several measures including relative humidity, temperature, and parameterizations of various cloud processes (ECMWF 2017). This means that this type of analysis is relatively unexplored with regards to snowfall forecasting. It should be noted that snowfall observations, given the lower-bounded nature of precipitation, do not follow a normal distribution, which was an assumed characteristic of observations in prior works. However, for the sake of testing this existing method which utilizes this assumption, the analysis that follows will make this same assumption. Of the utmost importance was whether the linear relationship between the ensemble variance and the error variance may be established and how well the linear regression performs for the January 2016 and January 2017 snowfall events.

CHAPTER III

RESULTS

Analysis of the 2010-2015 DJF EC EPS begins with the synoptic analysis, followed by the presentation of the ensemble skill. The mROC score and the other verification statistics for the entire period will be presented, followed by the analysis of the individual winter weather events. A discussion about the mROC score for the total period and the active winter weather periods is pursued. Analysis of the individual 3-day mROC averages and the resulting categories of predictability are examined, with discussion on how this information can be utilized. The second subsection of the results encompasses the linear variance calibration of the ensemble variance and the error variance. Details regarding the linear fits and discussion of the potential impact of the ensemble distribution, as well as the rank histograms, are considered. The results for the linear regression model are assessed through the differences noted between the modeled error variance and the actual error variance. These are placed in the context of anomalies in the synoptic weather pattern in addition to atmospheric forcing for vertical ascent and moisture transport.

EC EPS Snowfall Forecast Discrimination

The forecast discrimination suggests how accurately an ensemble distinguishes between an event (i.e. snowfall) and non-events (i.e. no snowfall). Note that this indicates nothing else about other observations of the forecast and solely represents hits and false alarms based on the defined threshold (any snowfall > 0). Also, the decision to establish the threshold for any snowfall removes possible errors related to SLR

conversions and assumes that any snowfall in the Southeast is impactful. A mROC above 0.5 indicates some measure of forecast discrimination above random chance. Similar to Hamill et al. 2008, where the uncalibrated EC EPS 5mm liquid equivalent reliability was measured as marginal, the DJF 2010-2015 snowfall forecasts were marginally reliable. As observed through Figure 2, the EC EPS can discriminate snowfall events over non-events skillfully through a 2-day lead time. Beyond that, the forecast system has a higher false alarm ratio than hit rate. This characteristic becomes more defined into the extended period beyond day 5 when the degradation of mROC accelerates.

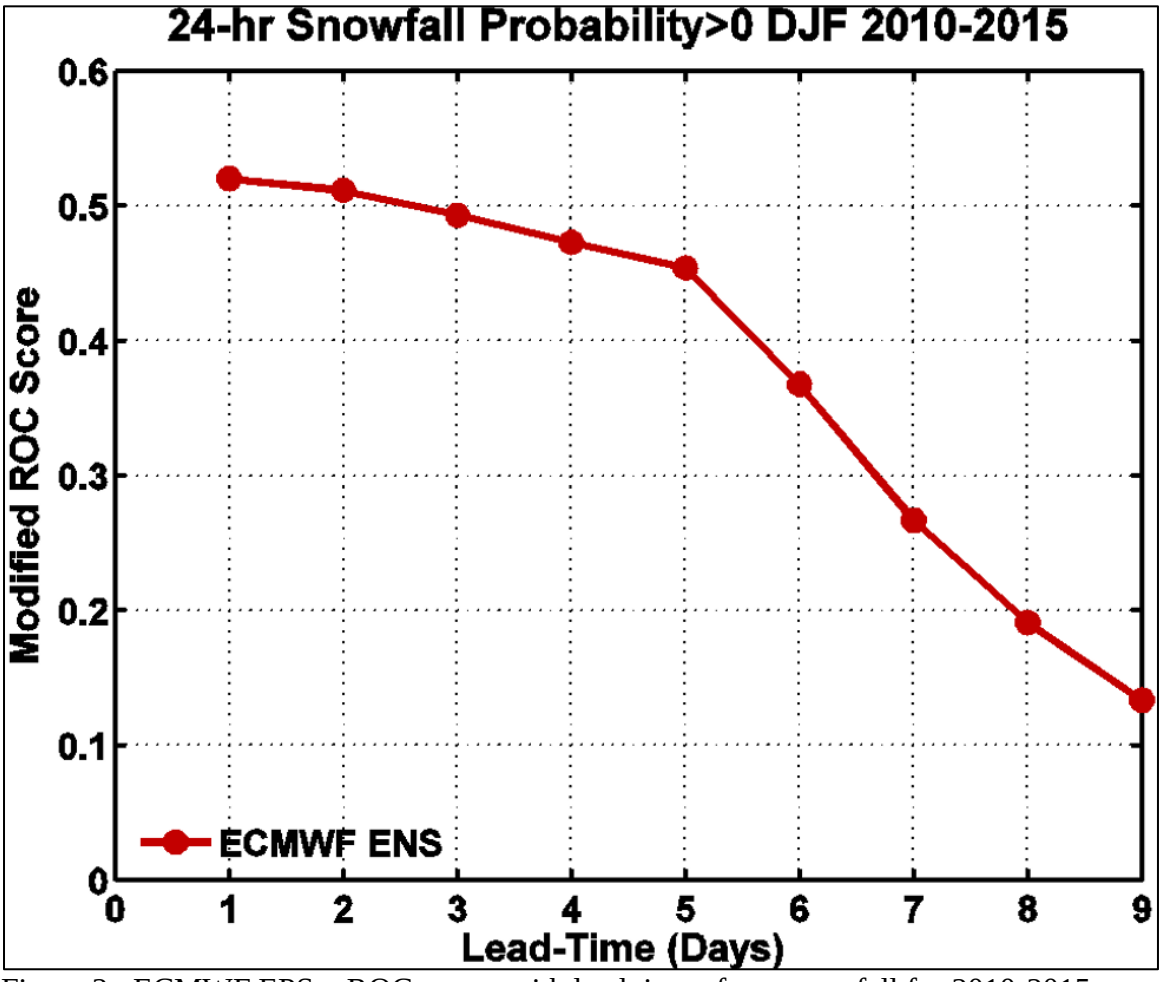

Figure 2: ECMWF EPS mROC scores with lead time of any snowfall for 2010-2015 DJF.

However, forecasters analyze and utilize the ensembles most during active weather patterns. Examining the mROC score over five winter seasons includes both active and quiet periods for Georgia. Several winter weather events were recorded throughout 2010-2011 and 2014-2015, but between these periods notable winter weather events were sparse (Table 1). However, the mROC score is a measure independent of the possible forecasts and reflects potential skill (Wilks 2014). Without prior knowledge of how the EC EPS performs from previous events, this only acts to inform to the consumer of how well the ensemble forecast system could distinguish an event from a non-event. Separating the active winter weather patterns in GA, the range of mROC scores for the selected 15 winter events indicate a pattern that similar to the 2010-2015 DJF period, albeit unreliable on average (Figure 3). Both display similar trends, where the degradation is most notable after Day 5. The range of mROC scores are large, with even a lead time of one day displaying high variability in potential skill for the selected winter weather events.

Although several winter weather events proved difficult to forecast, the range of mROC scores indicates that there are instances where highly skilled forecasts were made. However, in terms of consistency, it is not always the case as seen in Table 2. With the rarity of snowfall in the Southeastern United States, this presents a variety of communication challenges. For the operational meteorologist, these results suggest you cannot use the EC EPS to determine whether the winter weather event they are issuing forecasts for will be highly predictable or not. Without calibration, many EC EPS forecasts are unable to properly warn for any snowfall. Most of the EC EPS forecasts feature low discrimination, with an average 3-day mROC score below 0.5. However,

some displayed skillful forecasts within 72-hrs. Most events with higher discrimination were events that produced lighter snowfall amounts, with the 2011, Jan 9 event an exception (Table 2).

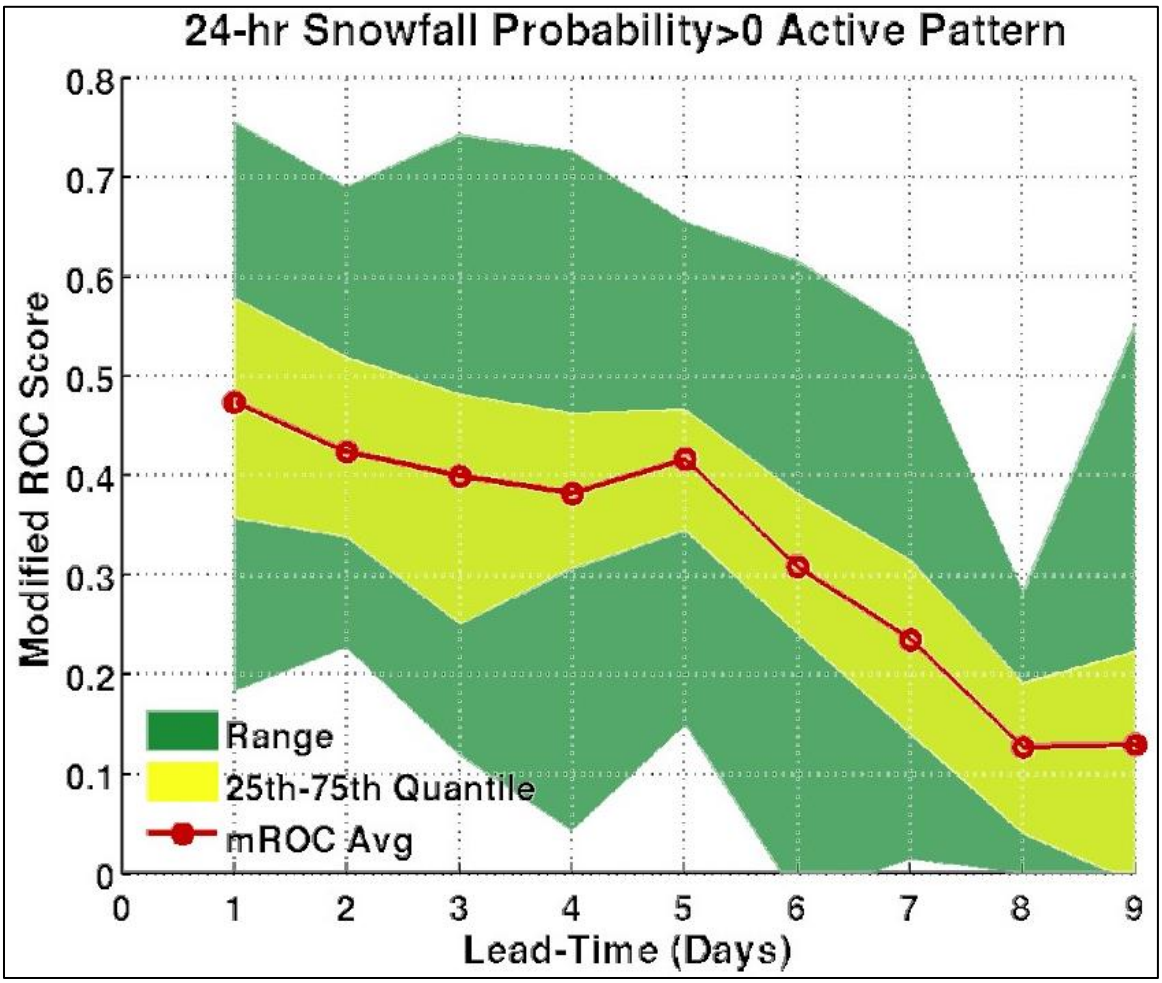

Figure 3: Same as Figure 2, but only considering active winter weather periods as selected in Table 1. The mROC average for the winter weather events is shown, and the range of extreme scores and the 25th-75th quantile are shaded.

Table 2: Predictability rating for the events listed in Table 1 designated into categories based on a 3-Day mROC average.

Date	Rating	3-Day mROC Avg.	1-Day Lead mROC Score	2-Day Lead mROC Score	3-Day Lead mROC Score
2010-01-08	Low	0.46	0.45	0.45	0.48
2010-02-12	Low	0.42	0.44	0.43	0.40
2010-02-16	Low	0.31	0.27	0.34	0.31
2010-12-13	Low	0.34	0.25	0.36	0.42
2010-12-16	Low	0.23	0.18	0.29	0.20
2010-12-26	High	0.68	0.76	0.69	0.59
2011-01-10	Marginal	0.52	0.57	0.54	0.46
2011-02-04	Low	0.40	0.66	0.42	0.12
2011-02-10	Low	0.39	0.60	0.34	0.23
2013-02-02	Low	0.25	0.35	0.23	0.18
2014-01-29	Low	0.46	0.57	0.34	0.47
2014-02-13	Low	0.40	0.49	0.39	0.33
2015-02-16	High	0.62	0.56	0.55	0.74
2015-02-21	Marginal	0.59	0.58	0.58	0.60
2015-02-25	Low	0.42	0.38	0.42	0.46

Snowfall Verification of ERA-Interim and NOHRSC Reports

As previously mentioned, the ERA-Interim reanalysis operates with the same IFS as the EC EPS. The observations bear a great deal of resemblance to the EC EPS forecasts. Overall, the NOHRSC data contains reports from a wide variety of sources, and is unrelated to the EC EPS forecasts. However, text reports were given in snowfall accumulation as opposed to snow water liquid equivalent and a SLR was applied. However, given that the ensemble variance is the predictor, SLR discrepancies or erroneous reports affects only the slope of the output. Rank histograms provide a mean of comparing the snowfall forecasts relative to observations as well as indicate EC EPS biases. Figure 4 contains the rank histograms for the 15 winter events as verified through ERA-Interim reanalysis and Figure 5 contains the same information verified through the NOHRSC reports. Both datasets reflect an under-dispersive EPS during active winter

weather patterns, with the rank histograms demonstrating that observations tend to lie at the extreme values of the ensemble members. Of note, the summation of the NOHRSC snowfall amounts are lower than the ERA-Interim totals, likely a result of the singular SLR value being applied across the entire region. However, for many of the events, the distributions appear visually similar across both datasets. The NOHRSC verification suggests greater under-dispersion than the ERA-Interim data, which display a few cases with a bias towards higher snowfall forecasts (Figure 4d, 4f, 4i, 4j).

Differences in the rank histogram distributions provide may be used to infer overprediction or underprediction of forecast error variance based on the ensemble variance. These inferences can be applied when assessing the individual distributions compared to the January 2016 and January 2017 events for verification of forecast error variance to observed values later in the results. Recall, events with histograms to the far left indicate over-forecast events with observations lower than most or all ensemble member forecasts (Figure 4a, 4c, 4d, 4f). Histograms containing observations greater than most all forecasts will be most prominent on the far right of the rank axis, indicating an under-forecast (Figure 4o, 5o). Most forecasts indicate an under-dispersive forecast, which demonstrates that prior forecaster concerns of under-dispersion are warranted for SE US snowfall events. These under-dispersive events are characterized by observations occurring at or outside the extremes of the ensemble member forecasts and indicates the ensemble members fails to capture the spectrum of possible outcomes. Many of the under-dispersive rank histograms also indicate a bias toward one of the forecast extremes. Only one particular event using the ERA-Interim reanalysis data indicates a potentially over-dispersive model with a bias towards over-forecasting (Figure 4), where the bin

frequency is greater towards the center and minimized over the extremes, indicating an ensemble forecast with a spectrum of snow forecasts that is too wide. For most snowfall events, the rank histograms are similar with a few exceptions. All the rank histograms characteristic of an over-forecasting ensemble family based on ERA-Interim verification were rendered under-dispersive based on the NOHRSC verification. Other notable differences include Figure 4i and 5i with observations that are higher than the NORHSC dataset and Figure 4m and 5m where the NOHRSC indicates a less under-dispersive forecast. These differences are important, because the slope of the linear regression model reflect the correction required to forecast the error variance based on the ensemble spread. The frequently under-dispersive ensemble forecast requires a greater slope to predict the error distribution.

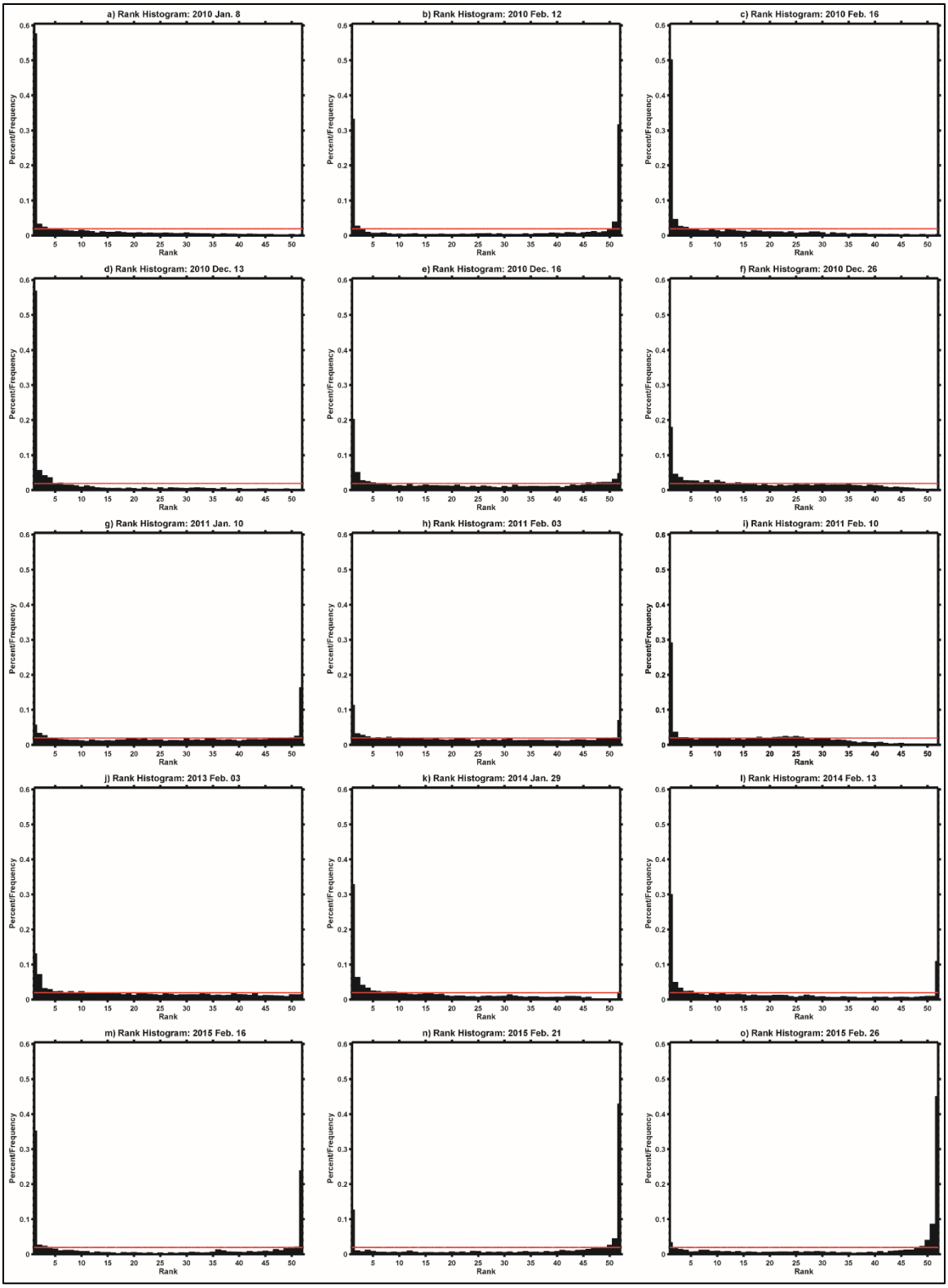

Figure 4: Rank histograms for the 15 winter weather events listed in Table 1 utilizing the ERA-Interim reanalysis data as verification. The red line indicates the relative frequency at which one would expect rank uniformity.

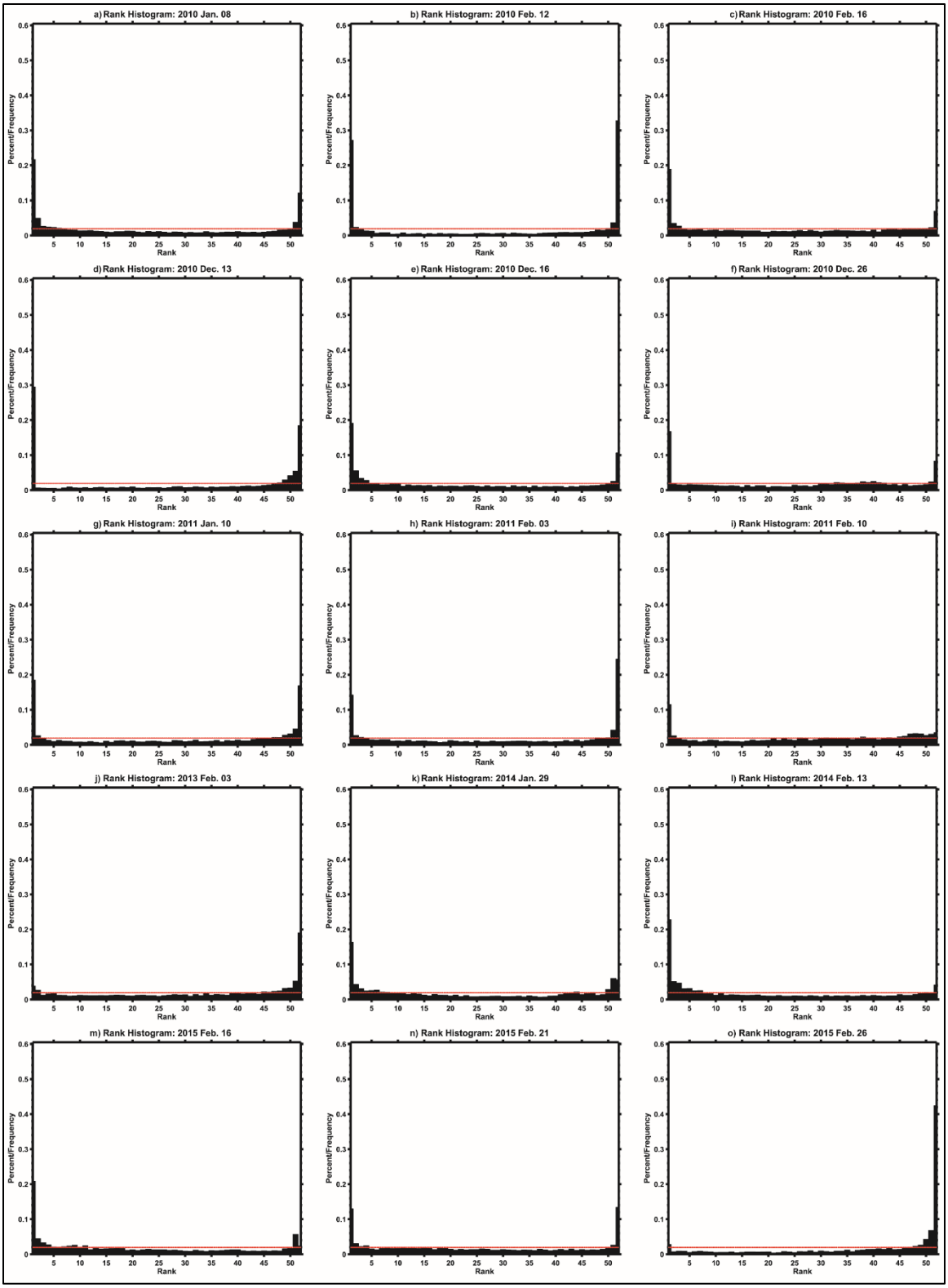

Figure 5: Same as Figure 4, but with the 15 winter weather events verified through the NOHRSC dataset.

The 22-23 January 2016 and 6-7 January 2017 winter weather events were different with regard to the prevailing synoptic weather pattern; however, both selected winter weather events contain many of the common elements noted in Konrad (1996) and Mote et al. (1997). Figure 6 and Figure 7 depict the 500 hPa GH contours with shaded anomalies, the 1000-850 hPa temperature anomalies and thickness contours associated with frozen precipitation, the specific humidity anomalies for the 1000-700 hPa layer, and the 850hPa GH contours and the wind speeds.

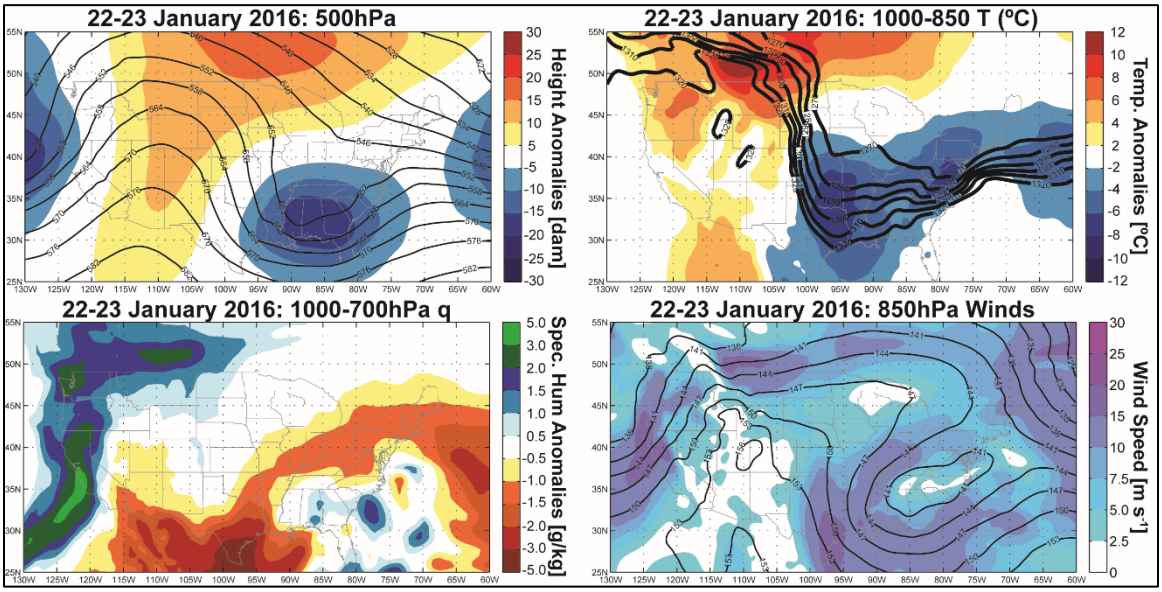

Figure 6: Top Left) 500hPa GH observed from 0000 UTC 22 January 2016 to 1200 UTC 23 January 2017 averaged observations over a 12-hr time step and contoured. Shaded 500 hPa anomalies based on a 1986-2015 climatology. Top Right) 1000-850 hPa layer mean temperature anomalies based on a 1986-2015 climatology and thickness values. Bottom Left) 1000-700 hPa mean layer specific humidity anomalies (g kg^{-1}) based on a 1986-2015 climatology. Bottom Right) 850 hPa wind speeds and GH contours.

Observations for the 22-23 January 2016 event reveal an anomalous ridge along the Rocky Mountains into Canada surrounded by an anomalous trough off the West Coast and a positively tilted trough along the Eastern United States at 500hPa. Negative 1000-850 hPa temperature anomalies of two to six degrees Celsius are common across

much of the SE US. Thicknesses values at the 1000-850 hPa layer less than 1300m are most favorable for snow, while values from 1300m to 1340m indicate mostly frozen precipitation is favored (Keeter et al. 1991). For this event, the 1300m is mostly above 35°N for the region of interest. The greatest moisture anomalies exist off the coastline. Based on the contours alone, the 850 hPa GH indicates a closed low and a warm front oriented SW to NE from North Georgia into Virginia, along which moisture was advected and where the greatest isentropic lift would have likely existed (Figure 6).

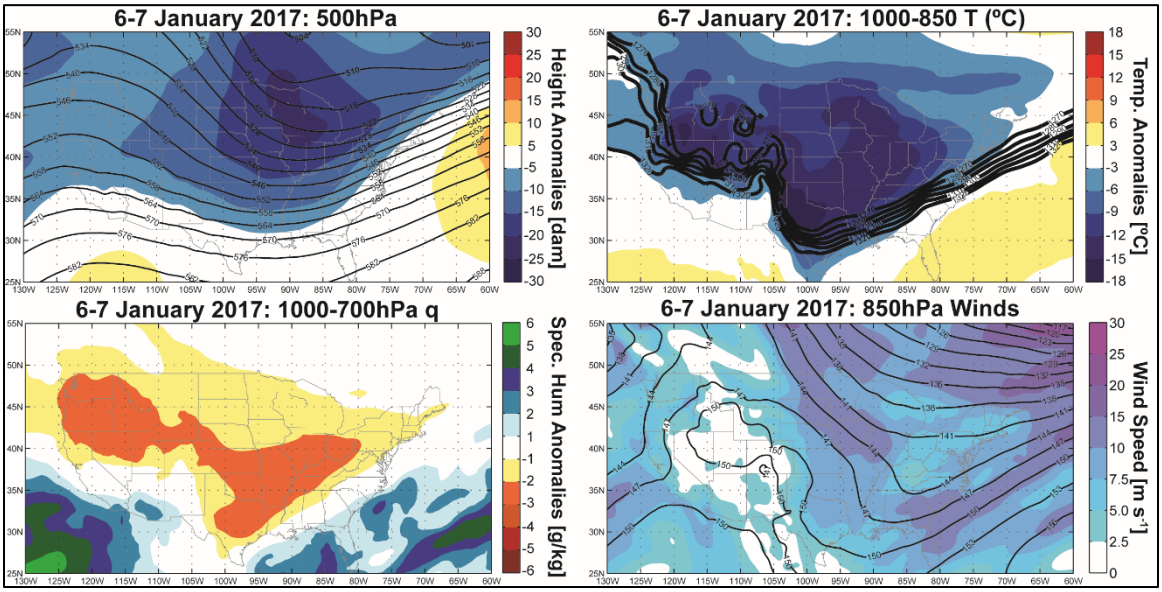

Figure 7: Top Left) 500hPa GH observed from 0000 UTC 6 January 2017 to 1200 UTC 7 January 2017 averaged observations over a 12-hr time step and contoured. Shaded 500 hPa anomalies based on a 1986-2015 climatology. Top Right) 1000-850 hPa layer mean temperature anomalies based on a 1986-2015 climatology and thickness values. Bottom Left) 1000-700 hPa mean layer specific humidity anomalies (g kg^{-1}) based on a 1986-2015 climatology. Bottom Right) 850 hPa wind speeds and GH contours.

Observations for the 6-7 January event reveal an anomalous trough prevailed across the United States. However, 500 hPa GH values indicate a ridge exists along the West Coast and a positively tilted trough along the Midwest. Negative temperature anomalies at the 1000-850 hPa layer were greater than the former event, but positioned well to the West of the SE US. Thus, the 1300m thickness orientation is more sharply

SW-NE, along Central Mississippi and Alabama into NE Georgia almost parallel to the Appalachian Mountain range. Anomalously dry air was present across much of the US, with greater moisture in the Gulf of Mexico. 850 hPa GH contours indicate a shallow shortwave not noticeable in the 500 hPa GH map that was embedded in the longwave trough. Geostrophic winds would support southwesterly winds for Georgia and the Carolinas, positioned well with anomalous low-level moisture, but also associated with warmer low-level air. As a result, snow would be favored along the Carolinas (Figure 7).

The 22-23 January 2016 generated a relatively large amount of snow in the 48-hr period from 1200 UTC January 22 to 12 UTC January 24. The maximum EC EPS mean forecast snow water liquid equivalent forecast up to 4cm located in Kentucky and Virginia for the period. Observed snowfall maxima occurred further north into West Virginia, with the forecast snow in Kentucky greater than observed. NOHRSC reports greater snowfall totals than the ERA-Interim reanalysis. Furthermore, snowfall along the Appalachian Mountain Range was also over-forecast (Figure 8).

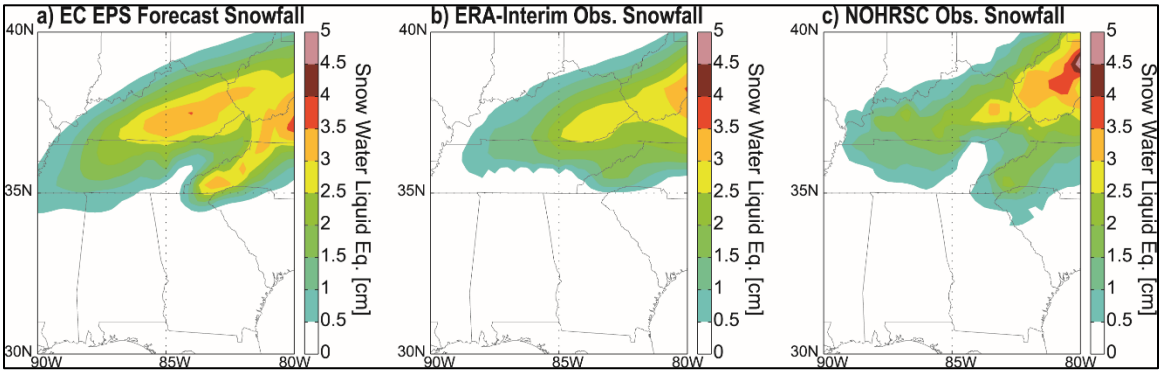

Figure 8: a) The ECMWF EPS mean forecast for snow water liquid equivalent (cm) for the 48-hr time period from 1200 UTC 22 January 2016 to 1200 UTC 24 January 2016 and b) the observed value for the same time period from ERA-Interim reanalysis and c) the observed value for the same time period from the NOHRSC reports.

The 6-7 January 2017 generated light snow mainly along the Appalachian Mountain Range. The ensemble mean forecast for snow water liquid equivalent between the 48-hr period from 1200 UTC 6 January 2017 to 1200 UTC 8 January 2017 time period provided an accurate forecast for the amount. However, verification from the ERA-Interim and NOHRSC indicate snowfall extended farther north. Notably, ERA-Interim extends light snowfall across Alabama, where this is absent from the NOHRSC reports (Figure 9). This may be a result of sleet and freezing rain still being designated as snow within ERA-Interim's microphysical schemes, which is reliant on an IFS cycle prior to cy41r1.

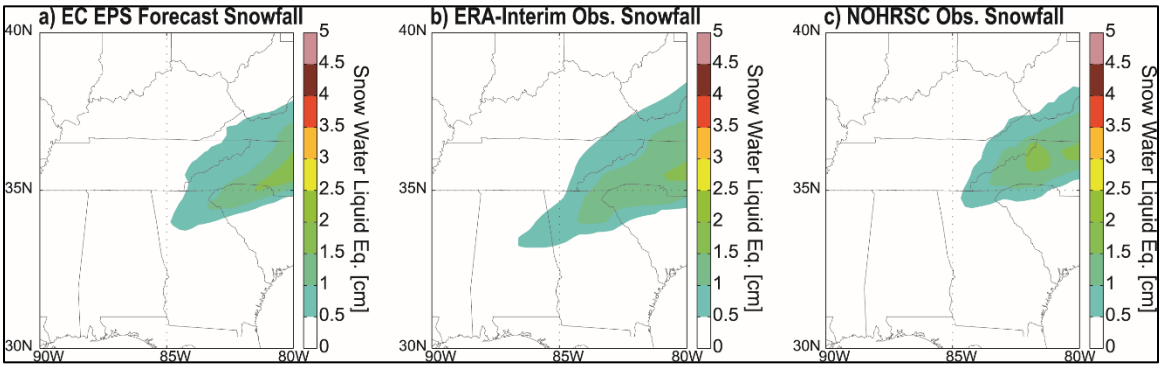

Figure 9: Same as Figure 8, but for the period 1200 UTC 6 January 2017 to 1200 UTC 8 January 2017.

Rank histograms for each event verify what can be seen visually (Figure 10). The 22-23 January 2016 event was over-forecast, especially in Kentucky. The rank histogram is similar to the selected 15 winter weather events from 2010-2015 and indicates the overestimated 48-hr snow water liquid equivalent. For the 6-7 January 2017 event, the rank histogram was under-dispersive comparatively. However, the smaller spatial extent means more ties must be settled at low snowfall thresholds, which may have allowed the rank histogram to appear more uniform. The NOHRSC verification almost suggests a

slight under-forecasting bias. However, both events appear to be less under-dispersive the 15 selected Southeastern winter weather events. Neither of the out-of-sample winter weather events had an mROC beyond the values of the selected winter weather events. The mROC score for the 22-23 January 2016 event was 0.36, indicating a less skillful forecast overall, while the mROC score for the 6-7 January 2016 event was 0.62. However, the smaller spatial extent operates in favor of the 6-7 January 2016 event.

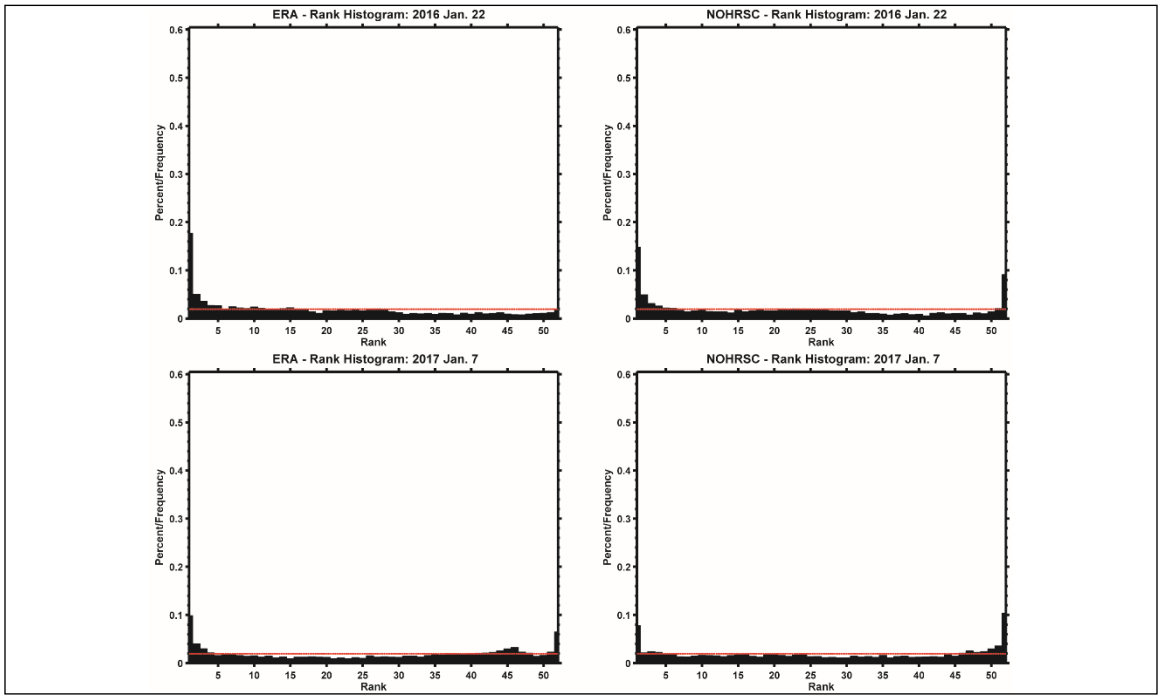

Figure 10: Rank histograms for winter weather events in January 2016 (top) and January 2017 (bottom) with the ERA-Interim reanalysis verification (left) and NOHRSC verification (right). The red line indicating the relative frequency at which rank uniformity would exist.

The Error Variance vs. Ensemble Variance Linear Model

Each set of observations output different error variances based upon the differences in how each of the datasets are constructed and for what purpose they were constructed. Figure 11 and Figure 12 show the scatter diagrams of the binned ensemble variance and the binned error variance for the ERA-Interim and NOHRSC observations

respectively. Each observation dataset display similar trends. With increasing lead time, the ensemble variance tends to increase in magnitude, mostly after the first day, while the error variance steadily decreases with lead time. For each lead-time and for each set of observations, the best fit lines and R^2 values were calculated to indicate how well the linear fit applies, which quantifies how much the linear variance calibration model of the ensemble variance explains the error variance. The training models were then applied to the out-of-sample snowfall events of 22-23 January 2016 and 6-7 January 2017 winter weather events.

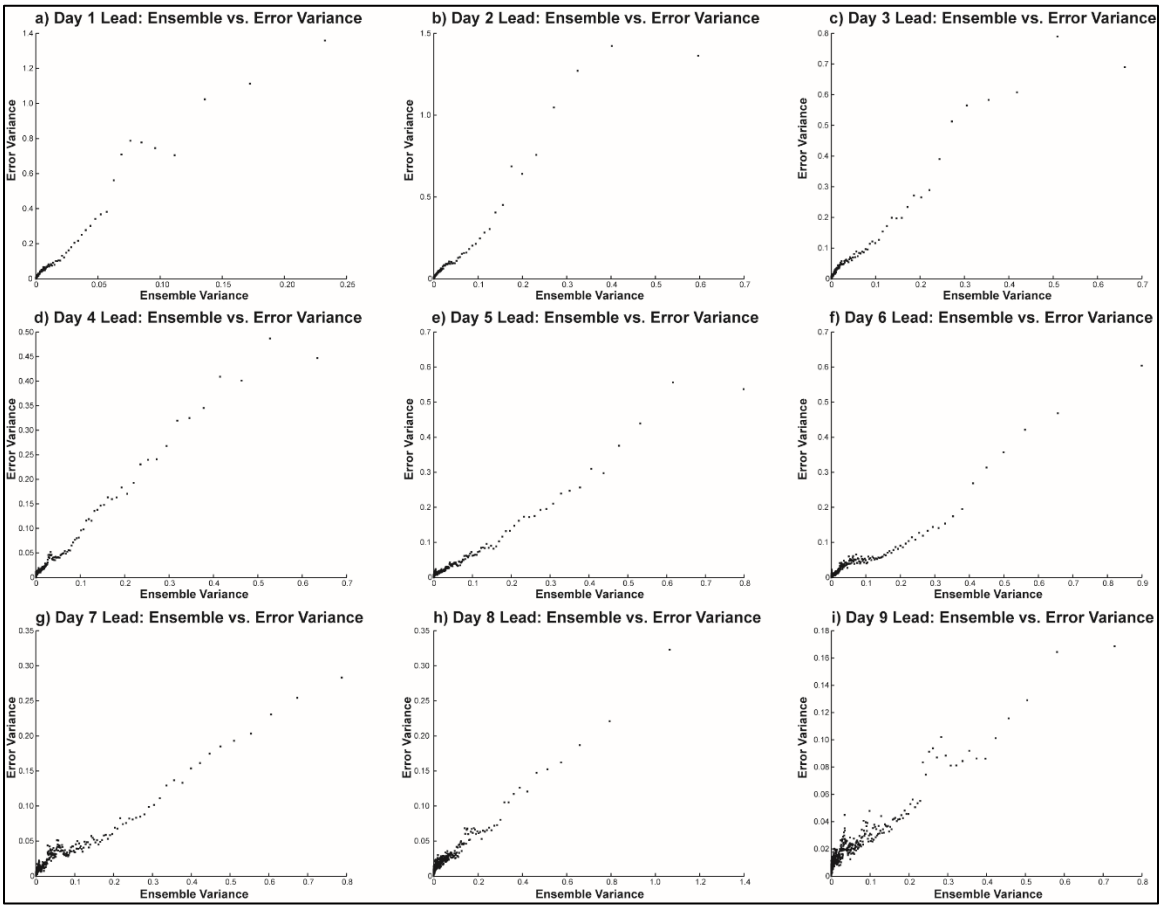

Figure 11: Scatter diagrams of the ensemble variance and error variance based on ERA-Interim for the 2010-2015 DJF period along the SE US. Note the differing scales and magnitude with lead-time.

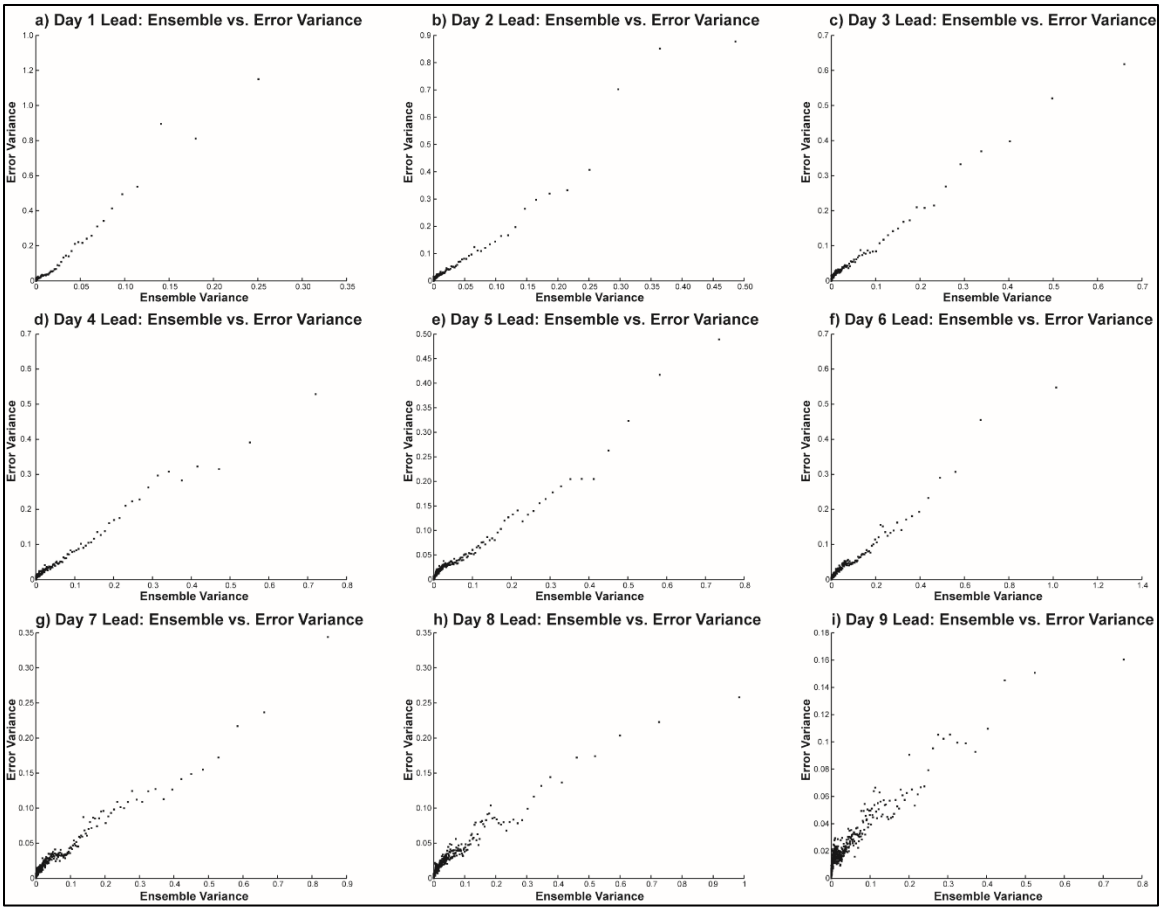

Figure 12: Same as Figure 11, but with error variance calculated using the NOHRSC snowfall database.

As stated prior, the summation of all snow water liquid equivalent within the NOHRSC observations was lower than the ERA-Interim reanalysis datasets. However, the linear regression model utilizes the ensemble variance as the predictand. Thus, the lower snow water liquid equivalent within the NOHRSC acts to reduce the resulting slope upon training. This linear variance calibration (e.g. Kolczynski et al. 2009) will vary depending on the observation dataset selected, but will ultimately retain the shape of the ensemble spread when contoured. However, a linear fit between the error variance and the ensemble variance is reasonably strong with high R^2 (> 0.90) values across all lead times. Particularly within the NOHRSC scatter diagram, the scatter points ascend

with a wavelike pattern, though this could indicate greater deviation than normal within the 1000 data points populating the bin (Figure 12). Not shown is the standard deviation for each bin, which typically increases as the ensemble variance increases, which is reasonable given the greater likelihood for errors in points with greater uncertainty. The intercept value of the best fit line steadily increases with lead-time, a result of the model trending towards climatological values. With the increasing ensemble variance and decreasing error variance with lead time, the slope of the best fit line decreases with time (Figure 13).

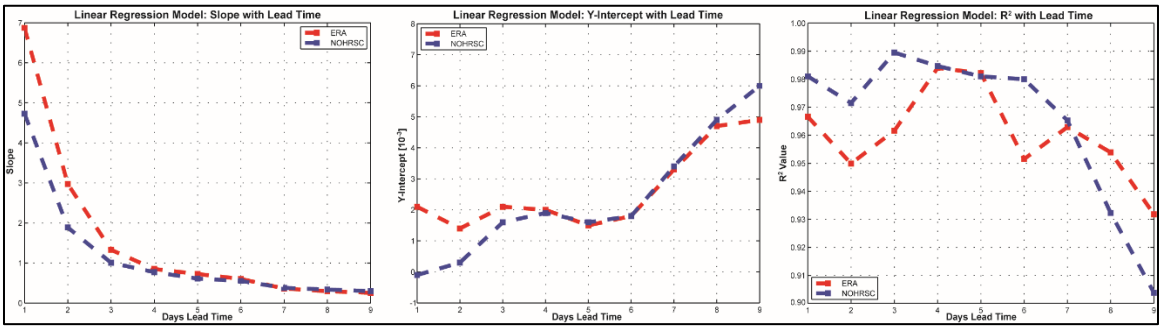

Figure 13: Left) Slope for the linear regression analysis of each dataset for nine-days lead time based on the 2010-2015 DJF period. Center) Y-intercept of the regression analysis. Right) R^2 value for each dataset measuring how well a linear model can fit the predictand with the predictor.

The results for the linear regression analysis contain a few notable differences from Kolczysnki et al. (2009). The trend with forecast time of the slope and the R^2 value are reversed. However, comparisons of forecasting the error variance based on ensemble variance between snow water liquid equivalent and wind forecast variables would be inappropriate. Furthermore, Kolczynski et al. (2009) utilized a mesoscale ensemble consisting of 10 members. Utilizing an idealized ensemble for further study of his linear variance calibration method, an increased number of ensemble members yielded an

increased slope, though not to the same degree as displayed in Figure 11. Once more, snow is prone to greater observed error given the sensitivity of the forecast variable.

Applying the model parameters to the 22-23 January 2016 and 6-7 January 2017, a forecast of the error variance was made utilizing the ensemble variance for the forecast 24-hrs in advance of the greatest snowfall. For the sake of brevity, the forecast model of the error variance of snow water liquid equivalent was shown for one day lead time. The 0000 UTC 22 January 2016 and the 0000 UTC 6 January 2017 were treated as the day one lead time since the succeeding 1200 UTC to 1200 UTC accumulations contained the maximum amount of snowfall for the event. Figure 14 displays the forecast ensemble variance for each event. Despite the fact that the 6-7 January 2017 event produced less snow than the 22-23 January 2016 event (Figure 9), the magnitude of the ensemble variance is approximately the same for each event. The ensemble variance acts as the predictand for the predictand, which is the error variance. The linear variance calibration assumes that the ensemble spread is an adequate predictor for where the greatest error occurs when generating the forecast, through which the slope of the regression analysis defines the ratio to which the ensemble variance must be made to allow a 1:1 relationship between the calibrated ensemble variance and the error variance. Incorrect error variance forecasts on the part of the linear variance calibration are the result of atypical levels of ensemble variability or greater forecast error than usual. Given the under-dispersive nature for many of the events, the calibration is adjusting to prevent underestimation of the error variance.

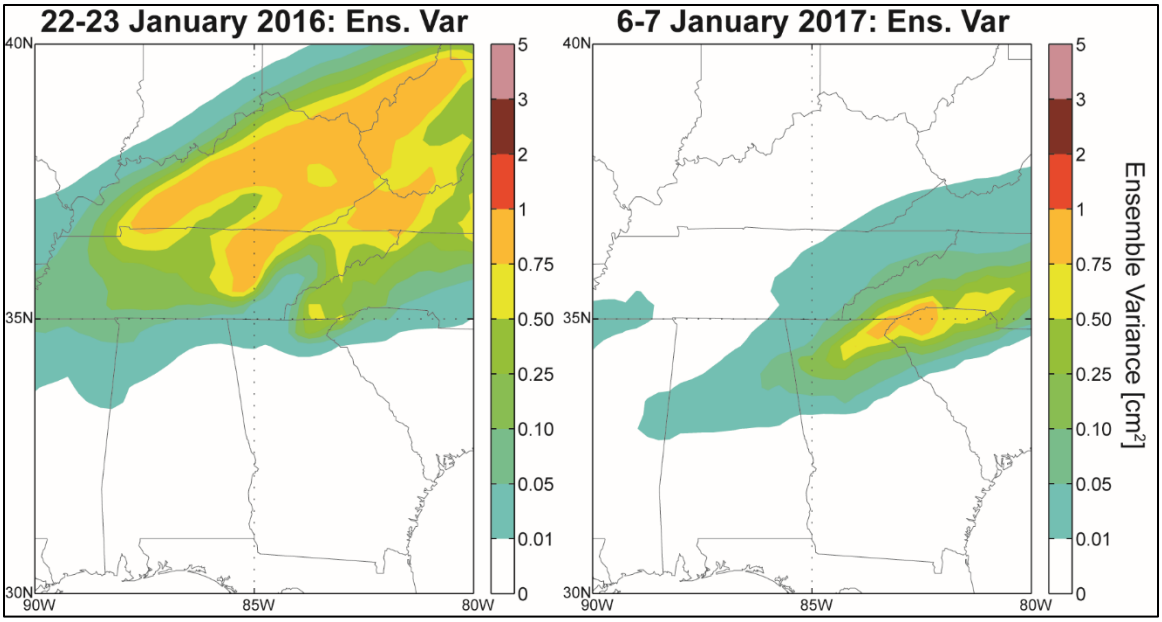

Figure 14: Left) Ensemble variance of the 24-hr snow water liquid equivalent (cm) for the time 1200 UTC 22 January 2016 to 1200 UTC 23 January 2017 from the 0000 UTC 22 January 2016 EC EPS initialization. Right) Ensemble variance of the 24-hr snow water liquid equivalent (cm) for the 0000 UTC 6 January 2017 EC EPS initialization.

Depicted in Figure 15 and Figure 16 are a comparison of the modeled and observed error variance for a one-day lead time as verified based on ERA-Interim reanalysis and NOHRSC for the 22-23 January 2016 and 6-7 January 2017 winter weather events, respectively. Results for the model forecast error variance for the 22-23 January 2016 and 6-7 January 2017 winter weather events indicate the greatest observed values of the error variance tend to lie near the local maxima of forecast snowfall of the EC EPS ensemble mean as opposed to the local maxima of ensemble forecast system's variability. This mainly applies to ERA-Interim, less so for NOHRSC given its independence from the EC EPS.

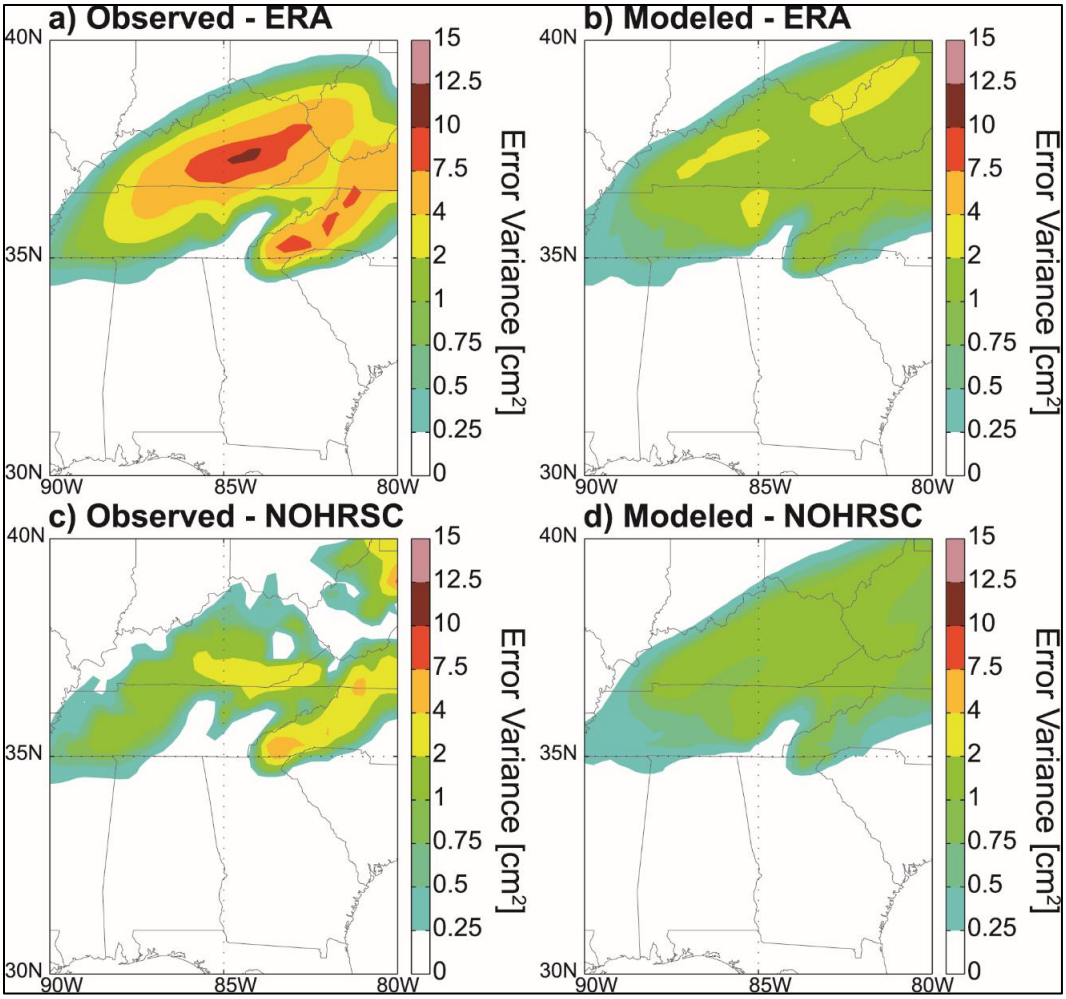

Figure 15: For the 24-hr snow water accumulation for the 1200 UTC 22 January 2016 to 1200 UTC 23 January 2016 period from the 0000 UTC 22 January 2016 EC EPS initialization, the observed (a, c) and modeled (b, d) error variance for ERA-Interim (Top) and the NOHRSC (Bottom) datasets.

The January 2016 event appears to be underestimated by both datasets, with the main shortcomings being the missed maxima in Kentucky and parallel to the Appalachian Mountains. The EC EPS forecast relatively higher amounts along the Appalachian Mountains, but the spread was low, resulting in the linear regression model to underestimate the error variance in that region (c.f., Figure 8 and Figure 14). The January 2017 event appears to overestimate the variance of error measured between observations and the ensemble mean. The modeled error variance indicates that some

ensembles generated snow as far west as Alabama, which was generally not observed. The mROC score for this event was 0.6210, indicating higher predictability than most events (Table 2), and it is reasonable to suggest that the observed error variance was less than normal, in addition to the more dispersive ensemble forecast (Figure 10). Common to both forecast linear regression models of the error variances was a northward bias for both events, particularly with ERA-Interim. Examination of the modeled error variance demonstrates that it retains the same shape as the ensemble variance, which extended or was maximized farther north than the forecast snowfall maxima (Figure 8 and Figure 9).

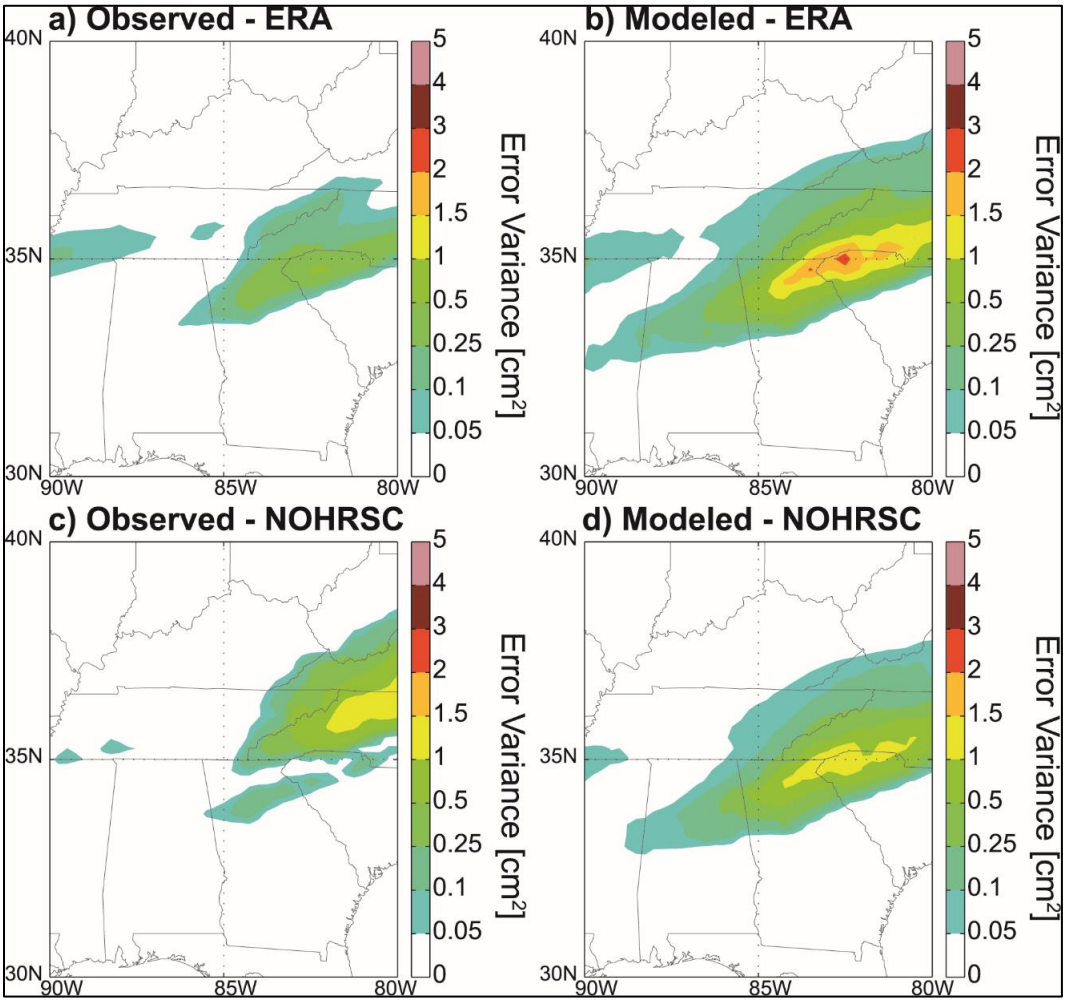

Figure 16: Same as Figure 15, but based on forecast snow water liquid equivalent for the 0000 UTC 6 January 2017 EC EPS initialization.

For each event and for each dataset, the ETS and bias score was also calculated. A threshold of 0.01 cm^2 was utilized first to get a sense as to how well the linear regression model represented the error variance of snow water liquid equivalent spatially. Table 3 lists the results for the January 2016 event, and Table 4 lists the results for the January 2017 event. Although the values for the forecast error variance for the January 2016 were underestimated, bias scores were greater than one for both the ERA-Interim verification and the NOHRSC verification. The ETS indicates the linear variance calibration was relatively skillful generating hits. However, bias scores are near or above one indicate that each dataset tended to overestimate the spatial extend of the error variance. Despite having a lower summed total accumulation, overestimation was greater in the NOHRSC generated linear variance calibration due to the differences in the observations from ERA-Interim. Fewer grid-points contained snowfall, allowing for the number of hits to be reduced as well.

Table 3: Equitable Threat Score and Bias Score calculations completed for the ERA-Interim dataset and the NOHRSC dataset. Results were calculated for the 0000 UTC 22 January 2016 EC EPS initialization of the 1200 UTC to 1200 UTC error variance in snow water liquid equivalent and the modeled error variance for each lead time.

0000 UTC 22 Jan 2016 Lead Time	ETS ERA	ETS NOHRSC	BS ERA	BS NOHRSC
Lead 1	0.83	0.68	1.09	1.17
Lead 2	0.66	0.55	1.19	1.29
Lead 3	0.65	0.45	1.18	1.38
Lead 4	0.80	0.66	1.10	1.15
Lead 5	0.56	0.41	1.29	1.46
Lead 6	0.67	0.51	1.20	1.27
Lead 7	0.16	0.27	2.79	1.82
Lead 8	0.29	0.37	1.96	1.51
Lead 9	0.49	0.51	1.35	0.86

The January 2017 event was less skillful in terms of the ETS. The differences between the ERA-Interim observations and the NOHRSC observations provided a boost to the scores of the NOHRSC generated linear variance calibrations. Furthermore, the bias for this event was higher than the January 2016 event. As mentioned earlier, the EC EPS members displayed greater spread than many snowfall events noted earlier (Figure 4, 5, and 10) and the EC EPS forecast skill was greater than normal. This likely impacted the modeled error variance to contain a greater number of false alarms within the dataset. However, the ETS is measured by the frequency of hits and is adjusted according to the relative number of hits. Based on both events, it is evident that the ensemble variance results in forecasting error variance tends to be too broad overall, but is capable of producing skillful forecasts beyond random chance.

Table 4: Same as Table 3, but beginning with the 0000 UTC 6 January 2017 EC EPS initialization.

0000 UTC 6 Jan 2017 Lead Time	ETS ERA	ETS NOHRSC	BS ERA	BS NOHRSC
Lead 1	0.31	0.42	2.27	1.84
Lead 2	0.30	0.45	2.13	1.28
Lead 3	0.40	0.53	1.67	1.44
Lead 4	0.25	0.31	2.29	1.64
Lead 5	0.51	0.55	1.40	1.22
Lead 6	0.24	0.37	1.45	1.17
Lead 7	-0.04	0.17	2.19	1.59
Lead 8	0.72	0.65	1.09	1.18
Lead 9	0.38	0.33	1.50	1.56

CHAPTER IV

CONCLUSIONS

The main purpose of this research is to evaluate the skill and behavior of an ensemble forecast system during Southeastern snowfall events and to use the information to respond to forecaster concerns regarding the ensemble forecast system utility. Chiefly, forecasters are wary of biased or under-dispersive ensemble forecasts and whether the ensemble mean adequately characterizes atmospheric processes better than a single deterministic model. This research analyzed the former issue, focusing on the skill of the EC EPS snow liquid water equivalent forecasts for the SE US. Among the forecasts for the 2010-2015 DJF period, the EC EPS' skill in discriminating snowfall of any kind was assessed using the mROC score. Forecasters have stated concern over applying a score than includes both active and quiet weather forecasts, so a closer examination of the selected winter weather patterns that occurred in and around the state of Georgia were also examined. Rank histograms were constructed for the active winter weather patterns to depict the EC EPS behavior. With a better understanding of the EC EPS behavior, information regarding the ensemble variance and the error variance for the 2010-2015 DJF period was employed to create a linear regression model, also referred to as a linear variance calibration model, to use forecast ensemble spread as a means to predict possible errors with respect to the ensemble mean. Two recent winter weather events that impacted Georgia on 22-23 January 2016 and 6-7 January 2017 were selected for the application of the linear variance calibration. For each event, the synoptic weather pattern was reviewed and compared with the findings of Mote et al. (1997). Each event

was analyzed in the same manner as the selected 15 active winter weather events. Comparisons of forecast error variance through the linear variance calibration model and the observed error variance generated through the ensemble mean were also completed. This was performed for two datasets of snowfall, ERA-Interim and NOHRSC. The ERA-Interim information of snowfall acts as an analysis field for the EC EPS. Due to model inter-dependency of ERA-Interim observations bore great resemblance with the EC EPS. The NOHRSC network of snow was included to address this concern and to determine how the performance of the linear variance calibrations was affected by the use of another set of observations. The ETS and the bias score were applied to these results to indicate the spatial skill of the linear variance calibration.

Analyses of the behavior of the EC EPS confirmed what many forecasters would expect for a Southeastern snowfall forecasts. For the 2010-2015 DJF period, the mROC score for any snowfall was greater than 0.5 up to a lead-time of two days. Even for the first two-days lead time, the score indicates only marginal skill above a random forecast. At a lead-time of three days, the mROC score falls below 0.5 and notably decreases further after day 5. Upon examination of the selected active winter weather events across the SE US, the mROC score for these events reflected the total period, even if none of the scores indicated a particularly useful forecast. However, the range of mROC scores and the 25^{th}-75^{th} percentile ranges indicated that several of these events were forecast successfully. Individually assessed, four of the fifteen events had a 3-day average mROC score above 0.5, and four others exhibited at least one skillful forecast within the 3-day period leading up to the event. This indicates that noteworthy winter weather events

were forecast with the approximately the same skill as winter weather events outside the fifteen selected events.

Research related to the distribution of observations compared to the individual ensemble forecasts indicates the EC EPS has a largely under-dispersive quality in active winter weather patterns. Few of the winter weather events that have impacted the SE US displayed a rank histogram that did not demonstrate this quality. A perfect ensemble forecast has a rank histogram with a uniform distribution, indicating that the ensemble equitably captures the possible range of forecast values. The under-dispersive quality seen for the selected active winter weather events were similar across both the ERA-Interim and NOHRSC observational datasets of snow water liquid equivalent. A couple of forecasters in the Evans et al. (2014) study refused to incorporate ensemble forecasts due to recognized under-dispersive qualities within the EPS.

This work also sought to use the ensemble variance to forecast the range of error from the ensemble mean forecast. Previous work completed by Roulston et al. (2005) and Kolczynski et al. (2009) focused on proposing an alternative method for using ensemble spread to forecast error variance. Assuming linearity exists between the second moments of a statistic, these values provided a better fit than previous methods. This method was applied in previous works to forecast variables related to various physical processes. Applications for the linear covariance model had not been applied to precipitation outputs, which represents a model variable that is heavily impacted by other physical schemes within the ensemble forecasting system. Scatter diagrams of the ensemble variance and the error variance for snow water liquid equivalent forecasts at various lead times were created for the 2010-2015 DJF period, treating each forecast

period independently of each other. The coefficient of determination of the best-fit line indicated that the data points were strongly suited for linear modeling with values > 0.9 at all lead times. This was true of each dataset, without consideration of the non-normal distribution of snowfall observations. The linear variance calibration model accounted for the under-dispersive behavior of the ensemble based on the slope of the best-fit line. This linear regression model had a slope well above 1 for the first two days of lead-time. Lead days 3 and 4 required little alteration, and for larger lags, the ensemble variance needed to be reduced to better fit the error variance. This is visualized through the slope of the linear regression model decreasing at greater lead times, which revealed the impact of increasing ensemble variance. The decrease in error variance with lead-time is an unexpected result from this analysis and should warrant additional consideration in future work. The most likely scenario is that the ensemble mean is scoring well compared to observations in the long-term given how small snow accumulation values are climatologically. Decreasing R^2 values with lead-time revealed that the ability for the linear variance calibration method to explain the error variance through the ensemble variance decreases. This contrasts with Kolczynski et al. (2009) where R^2 values increased with lead-time, but given their use of a mesoscale ensemble forecast system, the spin-up of the model likely caused this. However, the result of the changing slope and the changing R^2 values with lead-time confirms the importance of performing the linear regression analysis for each lead-time available.

Two recent events outside the 2010-2015 DJF period were analyzed: 22-23 January 2016 and 6-7 January 2017. The synoptic weather pattern for each event was distinct. The 22-23 January 2016 event featured a deep cyclone with relatively weak

negative cold air anomalies while the 6-7 January 2017 features a shallow cyclone with stronger negative cold air anomalies. Despite differences in strength, the common features noted by Konrad (1996) and (Mote et al. 1997) were present in both cases. Rank histograms for each event, along with the mROC scores, exhibit features quite similar to the selected winter weather events in the 2010-2015 DJF period.

Application of the linear regression models, separately constructed from the ERA-Interim and NOHRSC snowfall accumulations, to the 22-23 January 2016 and 6-7 January 2017 revealed several important issues related to the selection of the observation datasets. The spatial extent of the error variance model replicates the shape of the ensemble variance. While this is beneficial for regions with larger uncertainty, there are many regions for which only one or two ensembles produce precipitation values. Thus, the linear variance calibration of snow water liquid equivalent depicted error variance that was too broad in scope and overly dependent on the performance of the EC EPS, which has been shown to be widely varied for active Southeastern winter weather periods. As shown, the verification dataset selected for establishing the linear regression model will impact the slope of the model, and the verification network selected to verify the forecast error variance will cause different ETS and bias scores. How much one version of the linear regression model was favored over the other depended on how well the EC EPS forecast the event and how the present biases within the verification affected the observed error variance. Without a larger population of events to analyze, it is difficult to determine how significant these differences are. The issue is further compounded with continually updated EC EPS model cycles including the IFS cy41r1 update, which separated sleet and freezing rain into separate hydrometeor categories from

snowfall, limiting the extent to which new events can be added to the population. The inclusion of reforecasts would be another subject entirely and would also face difficulty measuring up to the benefits of a large EPS (Kolczysnki et al. 2011).

At the fundamental level, even an under-dispersive ensemble forecast can provide valuable information. This research has demonstrated that this method can be applied to forecast variables as sensitive as snowfall. The linear variance calibration method, when given a large amount of ensemble data, can adjust the values to remove the problems that under-dispersion causes when calculating the range of possible values from the ensemble mean. This provides a greater level of certainty for constructing forecast messages and probability forecasts for a range of values.

However, the linear variance calibration method requires additional adjustments to work most effectively for precipitation outputs. To produce optimal results would be computationally expensive, but even the simplest method can provide valuable details through adjustment of ensemble spread to account for an under-dispersive EPS. Reliance on the variance of one EPS is restrictive and leaves the model prone to the shortcomings of that EPS. Future efforts should seek to incorporate several ensemble prediction systems and evaluate the impact on the extent of the forecast error variance. Other remedies might include how the ensemble variance is included into the bins based on the ensemble mean. In other words, some method should separate those points for which four or five ensembles members forecast snow while the remaining ensemble members produce no precipitation. This would alleviate issues related to the broad spatial extent of the base forecast and may be essential for precipitation analysis, especially in convective environments. Despite the many things that need to be adjusted the linear variance

calibration may be as precise as one needs it to be. The method provides a useful resource for adjusting typical EPS characteristics. As any MOS forecast, the linear variance calibration method is prone to error for rapid changes or unusual circumstances, but takes advantage of the characteristics of the EPS to provide a better forecast than the raw model forecast.

REFERENCES

Andersson, E., and I. Tsonevsky, 2015: User Guide to ECMWF Forecast Products. *ECMWF*, 121.

Balsamo, G. S. Boussetta, P. Lopez, and L. Ferranti, 2010: Evaluation of ERA-Interim and ERA-Interim-GPCP-rescaled precipitation over the U.S.A, *ERA Report Series No. 5, ECMWF, Reading UK*.

Balsamo, G., C. Albergel, A. Beljaars, S. Boussetta, E. Brun, H. Cloke, D. Dee, E. Dutra, J. Muñoz-Sabater, E. Pappenberger, P. de Rosnay, T. Stockdale, and F. Vitart, 2015: ERA-Interim/Land: a global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.* **19**, 389-407.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.*, **66**, 603-626.

Bright, D. R. and P. A. Nutter, 2004: On the challenges of identifying the "best" ensemble member in operational forecasting. Preprints, *16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., J11.3.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168-189.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, *Q.J.R. Meteorol. Soc.*, **125**, 2887–2908.

Dee, D. P, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsama, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N Thépaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q.J.R. Meteorol. Soc.* **137**, 553-597.

DeVoir, G., 2004: High impact sub-advisory snow events: The need to effectively communicate the threat of short duration high intensity snowfall. *20th. Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc.

Du, J. and Z. Binbin, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Monthly Weather Review*, **139**, 3284-3303.

Ebert. E. E. and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrol.*, **239**, 179-202.

Ebert. E. E. 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, **129**, 2461-2480.

ECMWF, 2017: IFS documentation CY43R1. Accessed 24 July 2017. [Available online at http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation]

Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How Do Forecasters Utilize Output from a Convection-Permitting Ensemble Forecast System? Case Study of a High-Impact Precipitation Event. *Wea. Forecasting*, **29**, 466-486.

Forbes, R., I. Tsonevsky, T. Hewson, and M. Leutbecher, 2014: Towards predicting high-impact freezing rain events. *ECMWF Newsletter*, **141,** 15-21.

Gneiting, T. and A.E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310** (5746), 248-249.

Gneiting, T., 2014: Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum*, **719**.

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: stochastic ensemble results. *Monthly Weather Review*, **135**, 203-221.

Gurka, J. J., E. P. Auciello, A. F. Gigi, J. S. Waldstreicher, K. K. Keeter, S. Businger, and L. G. Lee, 1995: Winter weather forecasting throughout the Eastern United States. Part II: An operational perspective of cyclogenesis. *Wea. Forecasting*, **10**, 21-41.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-Meter temperatures. *Monthly Weather Review*, **136**, 2608-2619.

Haiden, T., M. Janousek, P. Bauer, J. Bidlot, M. Dahoui, L. Ferranti, F. Prates, D.S. Richardson, and F. Vitart, 2015: Evaluation of ECMWF forecasts including 2014-2015 upgrades. *ECMWF Technical Memorandum*, **765**.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, **126**, 711-724.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550-560.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS Ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, **136**, 2621-2632.

Harvey, O. L., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review*, **120**, 863-883.

Homan, J. and L. Uccellini, 1987: Winter forecast problems associated with light to moderate snow events in the Mid-Atlantic States on 14 and 22 February 1986. *Wea. Forecasting*, **2**, 206-228.

Houtemaker, P. L., 1993: Global and local skill forecasts. *Monthly Weather Review*, **121**, 1834-1846.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of Data Assimilations at ECMWF, *ECMWF Technical Memorandum*, **636**.

Keeter, K. K., S. Businger, L. G. Lee, and J. S. Waldstreicher, 1995: Winter weather forecasting throughout the Easter United States. Part III: The effects of topography and the variability of winter weather in the Carolinas and Virginia. *Wea. Forecasting*, **10**, 42-60.

Keil, C. and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Monthly Weather Review*, **135**, 3248-3259.

Kolczynski, W. C. Jr., D. R. Stauffer, S. E. Haupt, and A. Deng, 2009: Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *J. Applied. Meteorol.*, **48**, 2001-2021.

Kolczynski, W. C. Jr., D. R. Stauffer, and S. E. Haupt, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Monthly Weather Review*, **139**, 3954-3963.

Konrad, C. E., 1996: Relationships between the intensity of cold-air outbreaks and the evolution of synoptic and planetary-scale features over North America. *Monthly Weather Review*, **124**, 1067-1083.

Lackmann, G. M., K. K. Keeter, L. G. Lee, and M. B. Ek, 2002: Model representation of freezing and melting precipitation: Implications for winter weather forecasting. *Wea. Forecasting*, **17,** 1016-1033.

Mason, S. J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 714-725.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and Validation. *Q.J.R. Meteorol. Soc.*, **122**, 73-120

Mote, T. L., D. W. Gamble, S. J. Underwood, and M. L. Bentley, 1997: Synoptic-scale features common to heavy snowstorms in the southeast United States. *Wea. Forecasting*, **12(1)**, 5-23.

Nehrkorn, T., R. N. Hoffman, C. Grassotti, and J. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Q.J.R. Meteorol. Soc.*, **129**, 195-218.

Notaro, M., D. Lorenz, C. Hoving, and M. Schummer, 2014: Twenty-First-Century projections of snowfall and winter severity across Central-Eastern North America. *J. Clim.*, **27**, 6526-6550.

Novak, D. R., D. R. Bright, and M.J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069-1084.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. *ECMWF Technical Memorandum*, **188**.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher 2005: Representing model uncertainty in weather and climate prediction. *Annual Rev. Earth Planet Sci.*, **33**, 163-193.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and Weisheimer, 2009: Stochastic Parameterization and Model Uncertainty. *ECMWF Technical Memorandum*, **598**.

Peterson, A. T., M. Papeş, and Jorge Soberón, 2008: Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Eco. Modeling*, **213**, 63-72.

Roulston, M. S., 2005: A comparison of predictors of the error of weather forecasts. *Nonlinear Proc. Geoph.*, **12**, 1021-1032

Sivillo, J.K., J.E. Ahlquist, and Z. Toth 1997: An ensemble forecasting primer. *Wea. Forecasting*, **12**, 809-818.

Squires, M. F., J. H. Lawrimore, R. R. Heim, D. A. Robinson, M. R. Gerbush, and T.W. Estilow, 2014: The Regional Snowfall Index. *Bull. Amer. Meteor. Soc.*, **95**, 1835-1848.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74(12)**, 2317-2330.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379-398.

Whitaker, J. S., T.M. Hamill, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Monthly Weather Review*, **136**, 463-482.

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135**, 2379-2390.

Wilks, D. S., 2014: International Geophysics, 100: Statistical Methods in the Atmospheric Sciences. 2nd ed., Academic Press, 648.